(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
22 February 2001 (22.02.2001)

PCT

(10) International Publication Number
WO 01/12791 A1

(51) International Patent Classification[7]: C12N 9/02, 15/52

(21) International Application Number: PCT/US00/22038

(22) International Filing Date: 11 August 2000 (11.08.2000)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/148,850 12 August 1999 (12.08.1999) US

(71) Applicant *(for all designated States except US)*: **MAXYGEN, INC.** [US/US]; 515 Galveston Drive, Redwood City, CA 94063 (US).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: **SELIFONOV, Sergey A.** [RU/US]; 2240 Homestead Court, Los Altos, CA 94024 (US). **NEWMAN, Lisa, M.** [US/US]; 1137 B Reed Ave., Sunnyvale, CA 94086 (US).

(74) Agents: **QUINE, Jonathan, Alan**; The Law Offices of Jonathan Alan Quine, P.O. Box 458, Alameda, CA 94501 et al. (US).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published:
— *With international search report.*
— *Before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments.*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

WO 01/12791 A1

(54) Title: DNA SHUFFLING OF DIOXYGENASE GENES FOR PRODUCTION OF INDUSTRIAL CHEMICALS

(57) Abstract: This invention provides improved polypeptides, including dioxygenases, dehydrogenases, ligases and transferases that are useful for the biocatalytic synthesis of compounds such as α-hydroxycarboxylic acids, and aryl- and alkyl-hydroxy and carboxylic acid compounds. The polypeptides provided herein are improved in properties such as regioselectivity, enzymatic activity, stereospecificity, and the like. Methods for obtaining recombinant polynucleotides that encode these improved polypeptides are also provided, as are organisms that express the polypeptides and are thus useful for carrying out said biocatalytic syntheses. Also provided by the invention are methods for increasing said solvent resistance of organisms that are used in the synthetic methods.

DNA SHUFFLING OF DIOXYGENASE GENES FOR PRODUCTION OF INDUSTRIAL CHEMICALS

CROSS REFERENCE TO RELATED APPLICATIONS

This application claims priority from and the benefit of U.S. Provisional Application No. 60/148,850, filed August 12, 1999, pursuant to 35 USC 119(e) and any other applicable statute or rule.

This application is related to USSN 09/373,333, filed August 12, 1999, "DNA Shuffling to Produce Herbicide Selective Crops", Subramanian, Venkitswaran, et al., which is a non-provisional application of USSN 60/096,288, filed August 12, 1998, USSN 60/111,146, filed December 7, 1998, and USSN 60/112,746, filed December 17, 1998, and related to International Application No. PCT/US99/18394, filed August 12, 1999, "DNA Shuffling to Produce Herbicide Selective Crops."

This application is also related to USSN 09/373,928, filed August 12, 1999, "DNA Shuffling of Monooxygenase for Production of Industrial Chemicals," Affholter, Joseph A., et al., which is a non-provisional of USSN 60/096,271, filed August 12, 1998, and USSN 60/130,810, filed April 23, 1999, and related to International Application No. PCT/US99/18424, filed August 12, 1999, "DNA Shuffling of Monooxygenase for Production of Industrial Chemicals."

COPYRIGHT NOTICE PURSUANT TO 37 C.F.R. § 1.71(e)

FIELD OF THE INVENTION

This invention pertains to the shuffling of nucleic acids to achieve or enhance industrial production of chemicals by dioxygenase genes.

BACKGROUND OF THE INVENTION

Oxygen-containing organic chemicals such as organic acids, hydroxy carboxylic acids, alcohols, hydroxyaryls (*e.g.*, hydroxyaryl carboxylic acids, alkylphenols, *etc.*) and glycols are important classes of industrial chemicals. Typically, these products are generated by successive introduction of various chemical functional groups by oxidation, (trans)alkylation, reduction, desaturation and other reactions of inexpensive raw materials such as saturated and unsaturated hydrocarbons (alkanes, alkenes, etc) and simple aromatic compounds (benzene, ethyl benzene, cumene, naphthalene, styrene, toluene, xylenes, etc).

Dioxygenases (DOs) such as the arene dioxygenases (ADOs) typically catalyze limited oxidation of these basic chemical building blocks. While potentially interesting from an industrial standpoint, these enzymes typically do not exhibit sufficient turnover numbers and/or desired specificity or regioselectivity to make them usable as industrial catalysts.

Surprisingly, the present invention provides a general (broad utility) method for providing dioxygenase enzymes with higher activity and desired selectivity and specificity of the catalyzed reactions, thereby solving the problems outlined above, as well as providing a variety of other features which will be apparent upon review.

SUMMARY OF THE INVENTION

In the present invention, nucleic acid shuffling is used to generate new or improved dioxygenase ("DO") genes. These dioxygenase genes are used to provide dioxygenase enzymes, especially for industrial processes. These new or improved genes have surprisingly superior properties as compared to naturally occurring dioxygenase genes.

In the methods for obtaining dioxygenase genes, a plurality of parental forms (homologs) of a selected nucleic acid are recombined. The selected nucleic acid is derived either from one or more parental nucleic acid(s) which encodes a dioxygenase enzyme, or a fragment thereof, or from a parental nucleic acid which does not encode dioxygenase, but which is a candidate for nucleic acid shuffling to develop dioxygenase activity. The plurality of forms of the selected nucleic acid differ from each other in at least one (and typically two or more) nucleotides, and, upon recombination, provide a library of recombinant dioxygenase nucleic acids. The library can be an *in vitro* set of molecules, or present in cells, phage or the like. The library is screened to identify at least one

recombinant dioxygenase nucleic acid that exhibits distinct or improved dioxygenase activity compared to the parental nucleic acid or nucleic acids.

Many formats for libraries of nucleic acids are known in the art and each of these formats is generally applicable to the libraries of the present invention. For example, basic texts generally disclosing library formats of use in this invention include Sambrook *et al., Molecular Cloning, A Laboratory Manual* (2nd ed. 1989); Kriegler, *Gene Transfer and Expression: A Laboratory Manual* (1990); and *Current Protocols in Molecular Biology* (Ausubel *et al.*, eds., 1994)).

In a preferred embodiment, the starting DNA segments are first recombined by any of the formats described herein to generate a diverse library of recombinant DNA segments. Such a library can vary widely in size from having fewer than 10 to more than 10^5, 10^7, or 10^9 members. In general, the starting segments and the recombinant libraries generated include full-length coding sequences and any essential regulatory sequences, such as a promoter and polyadenylation sequence, required for expression. However, if this is not the case, the recombinant DNA segments in the library can be inserted into a common vector providing the missing sequences before performing screening/selection.

Although a variety of methods are optionally used, the library is typically generated by nucleic acid shuffling. One preferred method comprises initiating a polynucleotide amplification process on overlapping segments of a population of variant polynucleotides, e.g., allelic or species variants, at least one of which variant polynucleotides typically encodes a dioxygenase polypeptide. This amplification is typically carried out under conditions whereby one segment serves as a template for extension of another segment, to generate a population of recombinant polynucleotides. The recombinant polynucleotides are typically selected or screened for a desired property, e.g., improved dioxygenase activity. The overlapping segments are optionally produced by cleavage of the population of variant polynucleotides, e.g., by DNaseI digestion. Alternatively, the overlapping segments are produced by chemical synthesis or by amplification of the population of polynucleotides.

For example, shuffling optionally comprises recombining at least first and second forms of a nucleic acid that encodes a dioxygenase polypeptide, or fragment thereof, wherein the first and second forms differ from each other in two or more nucleotides. The library of recombinant polynucleotides is then typically expressed to obtain a library of recombinant polypeptides. In some embodiments, the method further comprises

recombining at least one recombinant polynucleotide that encodes a member of the library of recombinant polynucleotides that encodes a member of the library of recombinant dioxygenase polypeptides, which is the same or different from the first and second forms, to produce a further library of recombinant polynucleotides. The further library of recombinant polynucleotides is expressed to obtain a further library of recombinant dioxygenase polypeptides. These steps are optionally repeated until the further library of recombinant polynucleotides contains a desired number of different recombinant polynucleotides.

In another embodiment, shuffling comprises hybridizing at least two sets of nucleic acids, wherein a first set of nucleic acids comprises single-stranded nucleic acid templates and a second set of nucleic acids comprises at least one set of nucleic acid fragments. The method further comprises elongating, ligating, or both, sequence gaps between the hybridized nucleic acid fragments, to generate one or more substantially full-length chimeric nucleic acid sequences corresponding to the single-stranded nucleic acid templates. In other embodiments, the method optionally comprises denaturing the one or more substantially full-length chimeric nucleic acid sequences and the single-stranded nucleic acid templates; separating the substantially full-length chimeric nucleic acid sequences from the single-stranded nucleic acid templates; and fragmenting the separated substantially full-length chimeric nucleic acid sequences by nuclease digestion or physical fragmentation to provide chimeric nucleic acid fragments.

If the sequence recombination format employed is an *in vivo* format, the library of recombinant DNA segments generated already exists in a cell, which is usually the cell type in which expression of the enzyme with altered substrate specificity is desired. If sequence recombination is performed *in vitro*, the recombinant library is preferably introduced into the desired cell type before screening/selection. The members of the recombinant library can be linked to an episome or virus before introduction or can be introduced directly. In some embodiments of the invention, the library is amplified in a first host, and is then recovered from that host and introduced to a second host more amenable to expression, selection, or screening, or any other desirable parameter.

The manner in which the library is introduced into the cell type depends on the DNA-uptake characteristics of the cell type (*e.g.*, having viral receptors, being capable of conjugation, or being naturally competent). If the cell type is not susceptible to natural and chemical-induced competence, but is susceptible to electroporation, one preferably

employs electroporation. If the cell type is not susceptible to electroporation as well, one can employ biolistics. The biolistic PDS-1000 Gene Gun (Biorad, Hercules, Calif.) uses helium pressure to accelerate DNA-coated gold or tungsten microcarriers toward target cells. The process is applicable to a wide range of tissues, including plants, bacteria, fungi, algae, intact animal tissues, tissue culture cells, and animal embryos. One can employ electronic pulse delivery, which is essentially a mild electroporation format for live tissues in animals and patients. Zhao, *Advanced Drug Delivery Reviews* 17:257-262 (1995). Novel methods for making cells competent are described in co-pending application U.S. patent application Ser. No. 08/621,430, filed Mar. 25, 1996. After introduction of the library of recombinant DNA genes, the cells are optionally propagated to allow expression of genes to occur.

In selecting for dioxygenase activity, a candidate shuffled DNA can be tested for encoded dioxygenase activity in essentially any synthetic process. Common processes that can be screened include, for example, dihydroxylation of an aromatic ring, olefinic or polyenic alkene π-bond dihydroxylation, oxidation of methyl or methlylene groups attached to an aromatic ring, oxidation of methyl or methylene groups attached to a π-bond which is not a part of an aromatic system, sulfur heteroatom monooxygenation, desaturation of alkane groups attached to aromatic ring or non-aromatic π-bonds, oxidative elimination of halide (F, Cl, Br, I), nitrite, ammonia from halogen, nitro or amino substituted π-bonds (aromatic and/or olefinic), N- dealkylation and dearylation of alkylamino- and arylamino-substituted aromatic compounds, O-dealkylation of alkoxy- and aryloxy-substituted aromatic compounds, and conversion of an aromatic ring to a substituted conjugated diene.

Similarly, instead of, or in addition to, testing for an increase in dioxygenase specific activity, it is also desirable to screen for shuffled nucleic acids which produce higher levels of dioxygenase nucleic acid or enhanced or reduced recombinant dioxygenase polypeptide expression or stability encoded by the recombinant dioxygenase nucleic acid.

A variety of screening methods can be used to screen a library, depending on the dioxygenase activity for which the library is selected. By way of example, the library to be screened can be present in a population of cells. The library is selected by growing the cells in or on a medium comprising the chemical or compound to be oxidized or reduced and selecting for a detected physical difference between the oxidized or reduced form of the chemical or compound and the non-oxidized or reduced form of the chemical or compound, either in the cell, or the extracellular medium.

Iterative selection for dioxygenase nucleic acids is also a feature of the invention. In these methods, a selected nucleic acid identified as encoding dioxygenase activity can be shuffled, either with the parental nucleic acids, or with other nucleic acids (*e.g.*, mutated forms of the selected nucleic acid) to produce a second shuffled library. The second shuffled library is then selected for one or more form of dioxygenase activity, which can be the same or different than the dioxygenase activity previously selected. This process can be iteratively repeated as many times as desired, until a nucleic acid with optimized properties is obtained. If desired, any dioxygenase nucleic acid identified by any of the methods herein can be cloned and, optionally, expressed.

The invention also provides methods of increasing dioxygenase activity by whole genome shuffling. In these methods, a plurality of genomic nucleic acids are shuffled in a cell (in whole cell shuffling, entire genomes are shuffled, rather than specific sequences). The resulting shuffled nucleic acids are selected for one or more dioxygenase traits. The genomic nucleic acids can be from a species or strain different from the cell in which dioxygenase activity is desired. Similarly, the shuffling reaction can be performed in cells using genomic DNA from the same or different species, or strains. Strains or enzymes exhibiting enhanced or modified DO activity can be identified.

The distinct or improved dioxygenase activity encoded by a nucleic acid identified after shuffling can encode for one or more of altered properties selected from a variety of properties of practical interest.

Of particular relevance to the practical use of DOs in making industrial chemicals are several types of improvements in catalytic properties and in the chemistry of the catalyzed reactions. Preferred catalytic properties of DOs that can be altered in a variety of combinations using nucleic acid shuffling are as follows:

1. Rate of reaction which can be catalyzed by enzyme.

Changing the rate of oxidative reactions (turnover numbers, V_{max} or other related parameters indicative of reaction rate) for a given substrate, or for a set of substrates, is often useful as wild-type dioxygenases accept substrates of varying structure, but the reaction is often so slow as to be impractical for preparative use.

2. Specificity of oxidation.

Altering the specificity of oxidation is preferred when the oxidation of a compound of interest occurs in a mixture of structurally related compounds, some of which may also serve as DO substrates (*e.g.* isomers of xylene and other alkylbenzenes). Evolving DOs by nucleic acid shuffling for greater specificity towards a particular compound of interest provides a means for using these enzymes in, for example, the reactive separation of compound mixtures, thus allowing for downstream separation of undesired substrates and their alternative uses.

3. Regioselectivity of oxidation.

Often, DOs form multiple or alternative products from a substrate when alternative substrate binding is possible and two or more sites for introducing oxygen are present. For the purpose of making industrial chemicals, high or otherwise altered selectivity of DO reaction is often preferred. Other properties roughly falling into the category of regioselectivity of reactions catalyzed by DOs include, for example, absolute configuration of chiral products and their enantiomeric purity (e.g., chiral hydroxylation), where chirality of products is possible.

4. Mode of action.

In respect to different substrates, or even to a single compound, DOs can catalyze different reactions including, but not limited to:

a) monooxygenation of sulfur atoms in various thioethers;

b) O- and N-dealkylation of appropriately substituted arenes;

c) oxidative dehalogenations and denitrations of halogenated and nitrated arenes;

d) monoxygenation (*e.g.*, monohydoxylation) of "benzylic" (*i.e.*, attached to a benzene or other aromatic ring) carbon atoms, whether methylene or methyl groups;

e) monooxygenation (*e.g.*, monohydoxylation) of allylic (*i.e.*, attached to a non-aromatic π-bond) carbon atoms, whether methylene and methyl groups;

f) desaturation reactions of alkyl or cycloalkyl carbon fragments attached to an aromatic ring (*e.g.* formation of styrene from ethylbenzene and indene from indan);

For the purpose of using arene dioxygenases ("ADO"s) in making industrial chemicals, it is often desired to control or change mode of action of wild-type (natural)

dioxygenases, e.g. dioxygenase reaction versus monooxygenase reaction, where two or more functional groups amenable to these reactions are present in the substrate's structure.

Among other properties of interest are those which may not be directly associated with the catalytic mechanism; however these physical and other general properties can make a profound impact on biocatalyst performance in a practical setting. These properties include, for example: an increase in the range of dioxygenase substrates which the distinct or improved polypeptide operates on, an increased expression level of a polypeptide encoded by the nucleic acid, a decrease in susceptibility of a polypeptide encoded by the nucleic acid to protease cleavage, a decrease in susceptibility of a polypeptide encoded by the nucleic acid to high or low pH levels, a decrease in susceptibility of the protein encoded by the nucleic acid to high or low temperatures, an optimization of nucleic acid codon usage for effective expression of ADO polypeptides in a particular host cell, a reduction in the sensitivity of the ADO polypeptides and/or an organism expressing the polypeptide to inactivation by organic solvents, a decrease in ADO inactivation or inhibition by substrates, products or reactive intermediates arising from ADO reaction or from other metabolic reactions of a host cell, and a decrease in toxicity to a host cell of a polypeptide encoded by the selected nucleic acid

The selected nucleic acids to be shuffled can be from any of a variety of sources, including synthetic or cloned DNAs. Exemplary targets for recombination include nucleic acids encoding arene dioxygenases and the like. Typically, shuffled nucleic acids are cloned into expression vectors to achieve desired expression levels.

One feature of the invention is production of libraries and shuffling mixtures for use in the methods as set forth above. For example, a phage display library comprising shuffled forms of a nucleic acid is provided. Similarly, a shuffling mixture comprising at least three homologous DNAs, each of which is derived from a nucleic acid encoding a polypeptide or polypeptide fragment is provided. These polypeptides can be, for example, arene dioxygenases, and the like.

Isolated nucleic acids identified by selection of the libraries in the methods above are also a feature of the invention.

Also provided are biotransformative methods using the dioxygenases of the invention for preparing diverse oxidized organic species. Representative species include vicinal diols, hydroxylated aromatic carboxylic acids, hydroxy alkylarene, α-hydroxcarboxylic acids and the like. Methods for preparing adducts of oxidized organic

species are also provided. To aid in practicing the biotransformative methods of the invention, there are also provided improved polypeptides and host organisms expressing these polypeptides.

Additional objects and advantages of this invention will be apparent to those of skill in the art from the detailed description that follows.

BRIEF DESCRIPTION OF THE FIGURES

Figure 1. Schematic showing the initial dioxygenation of an arene π-bond of a trialkylbenzene to produce a diol using a dioxygenase and the subsequent dehydration of this diol to a hydroxy trialkylbenzene.

Figure 2. Schematic showing the conversion of a monohydroxy arene to a dihydroxy arene using a dioxygenase.

Figure 3. Schematic showing esterification and de-esterification strategies using transferases, esterases and chemical dehydration.

Figure 4. Schematic showing the initial dioxygenation of an arene π-bond of a dialkylbenzene to produce a diol using a dioxygenase and the subsequent dehydration of this diol to a hydroxy dialkylbenzene.

Figure 5. Schematic showing the oxygenation of an alkylarene to the corresponding arene carboxylic acid, the subsequent oxidation of an arene π-bond to the corresponding diol and the alkylation and/or acylation of the carboxylic acid and/or hydroxyl moieties of the diol.

Figure 6. Schematic of the selective oxidation of one alkyl group of a dialkylarene to the corresponding arene carboxylic acid, the subsequent oxidation of an arene π-bond to the corresponding diol, the dehydration of the diol and the alkylation and/or acylation of the carboxylic acid and/or hydroxyl moieties of the diol.

Figure 7. Schematic showing the oxygenation of an alkylarene to the corresponding arenealkyl carboxylic acid, the subsequent oxidation of an arene π-bond to the corresponding diol and the alkylation and/or acylation of the carboxylic acid and/or hydroxyl moieties of the diol.

Figure 8. Schematic showing coumarin and coumarin derivatives.

Figure 9. Schematic showing synthetic routes to cinnamic acid, coumarin and derivatives of these compounds.

Figure 10. Schematic showing the oxygenation of a π-bond of a polycyclic arene to the corresponding diol, the opening of the diol functionalized ring and the conversion of the ring opened product into a lactone.

Figure 11. Structures of exemplary feedstock olefinic compounds and structures of α-hydroxy carboxylic acids produced from these feedstocks by dioxygenase action.

Figure 12. Enzymatic reaction schemes for multistep biochemical transformations of olefins to AHAs.

Figure 13. Enzymatic reaction schemes for converting free AHAs to ester derivatives.

Figures 14A- 14R. Table of presently preferred substrates and oxygenations.

Figure 15. Enzymatic reaction schemes for multistep biochemical transformations of olefins to AHAs.

Figure 16. Schematic illustrating crossovers for clones obtained from dioxygenase shuffling.

The absolute configuration of the chiral centers is not indicated in the above Figures. The chiral centers of the chiral compounds can be R, S, or a mixture of these configurations.

DETAILED DESCRIPTION OF THE INVENTION

Abbreviations

"AHA" refers to an α-hydroxycarboxylic acid.

"HCA" refers to a hydroxylated aromatic carboxylic acid.

"ADO" refers to an arene dioxygenase.

"DO" refers to a dioxygenase.

Definitions

Unless clearly indicated to the contrary, the following definitions supplement definitions of terms known in the art.

The term "shuffling" is used herein to indicate recombination between non-identical sequences, in some embodiments shuffling may include crossover via homologous

recombination or via non-homologous recombination, such as via cre/lox and/or flp/frt systems. Shuffling can be carried out by employing a variety of different formats, including for example, *in vitro* and *in vivo* shuffling formats, in silico shuffling formats, shuffling formats that utilize either double-stranded or single-stranded templates, primer based shuffling formats, nucleic acid fragmentation-based shuffling formats, and oligonucleotide-mediated shuffling formats, all of which are based on recombination events between non-identical sequences and are described in more detail or referenced herein below, as well as other similar recombination-based formats.

A "recombinant" nucleic acid is a nucleic acid produced by recombination between two or more nucleic acids, or any nucleic acid made by an *in vitro* or artificial process. The term "recombinant" when used with reference to a cell indicates that the cell includes (and optionally replicates) a heterologous nucleic acid, or expresses a peptide or protein encoded by a heterologous nucleic acid. Recombinant cells can contain genes that are not found within the native (non-recombinant) form of the cell. Recombinant cells can also contain genes found in the native form of the cell where the genes are modified and re-introduced into the cell by artificial means. The term also encompasses cells that contain a nucleic acid endogenous to the cell that has been artificially modified without removing the nucleic acid from the cell; such modifications include those obtained by gene replacement, site-specific mutation, and related techniques.

A "recombinant dioxygenase nucleic acid" is a recombinant nucleic acid encoding a protein or RNA which confers dioxygenase activity to a cell when the nucleic acid is expressed in the cell.

A "plurality of forms" of a selected nucleic acid refers to a plurality of homologs of the nucleic acid. The homologs can be from naturally occurring homologs (*e.g.*, two or more homologous genes) or by artificial synthesis of one or more nucleic acids having related sequences, or by modification of one or more nucleic acid to produce related nucleic acids. Nucleic acids are homologous when they are derived, naturally or artificially, from a common ancestor sequence. During natural evolution, this occurs when two or more descendent sequences diverge from a parent sequence over time, *i.e.*, due to mutation and natural selection. Under artificial conditions, divergence occurs, *e.g.*, in one of two ways. First, a given sequence can be artificially recombined with another sequence, as occurs, *e.g.*, during typical cloning, to produce a descendent nucleic acid. Alternatively, a nucleic acid

can be synthesized *de novo;* by synthesizing a nucleic acid which varies in sequence from a given parental nucleic acid sequence.

When there is no explicit knowledge about the ancestry of two nucleic acids, homology is typically inferred by sequence comparison between two sequences. Where two nucleic acid sequences show sequence similarity it is inferred that the two nucleic acids share a common ancestor. The precise level of sequence similarity required to establish homology varies in the art depending on a variety of factors. For purposes of this disclosure, two sequences are considered homologous where they share sufficient sequence identity to allow recombination to occur between two nucleic acid molecules. Typically, nucleic acids require regions of close similarity spaced roughly the same distance apart to permit recombination to occur.

The terms "identical" or percent "identity," in the context of two or more nucleic acid or polypeptide sequences, refer to two or more sequences or subsequences that are the same or have a specified percentage of amino acid residues or nucleotides that are the same, when compared and aligned for maximum correspondence, as measured using one of the sequence comparison algorithms described below (or other algorithms available to persons of skill) or by visual inspection.

The phrase "substantially identical," in the context of two nucleic acids or polypeptides (*e.g.*, DNAs encoding a dioxygenase, or the amino acid sequence of the dioxygenase) refers to two or more sequences or subsequences that have at least about 60%, preferably 80%, most preferably 90-95% nucleotide or amino acid residue identity, when compared and aligned for maximum correspondence, as measured using one of the following sequence comparison algorithms or by visual inspection. Such "substantially identical" sequences are typically considered to be homologous. Preferably, the "substantial identity" exists over a region of the sequences that is at least about 50 residues in length, more preferably over a region of at least about 100 residues, and most preferably the sequences are substantially identical over at least about 150 residues, or over the full length of the two sequences to be compared.

For sequence comparison and homology determination, typically one sequence acts as a reference sequence to which test sequences are compared. When using a sequence comparison algorithm, test and reference sequences are input into a computer, subsequence coordinates are designated, if necessary, and sequence algorithm program parameters are designated. The sequence comparison algorithm then calculates the percent

sequence identity for the test sequence(s) relative to the reference sequence, based on the designated program parameters.

Optimal alignment of sequences for comparison can be conducted, *e.g.*, by the local homology algorithm of Smith & Waterman, *Adv. Appl. Math.* **2**:482 (1981), by the homology alignment algorithm of Needleman & Wunsch, *J. Mol. Biol.* **48**:443 (1970), by the search for similarity method of Pearson & Lipman, *Proc. Nat'l. Acad. Sci. USA* **85**:2444 (1988), by computerized implementations of these algorithms (GAP, BESTFIT, FASTA, and TFASTA in the Wisconsin Genetics Software Package, Genetics Computer Group, 575 Science Dr., Madison, WI), or by visual inspection (*see generally*, Ausubel *et al.*, *infra*).

One example of an algorithm that is suitable for determining percent sequence identity and sequence similarity is the BLAST algorithm, which is described in Altschul *et al.*, *J. Mol. Biol.* **215**:403-410 (1990). Software for performing BLAST analyses is publicly available through the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/). This algorithm involves first identifying high scoring sequence pairs (HSPs) by identifying short words of length W in the query sequence, which either match or satisfy some positive-valued threshold score T when aligned with a word of the same length in a database sequence. T is referred to as the neighborhood word score threshold (Altschul *et al.*, *supra*). These initial neighborhood word hits act as seeds for initiating searches to find longer HSPs containing them. The word hits are then extended in both directions along each sequence for as far as the cumulative alignment score can be increased. Cumulative scores are calculated using, for nucleotide sequences, the parameters M (reward score for a pair of matching residues; always > 0) and N (penalty score for mismatching residues; always < 0). For amino acid sequences, a scoring matrix is used to calculate the cumulative score. Extension of the word hits in each direction are halted when: the cumulative alignment score falls off by the quantity X from its maximum achieved value; the cumulative score goes to zero or below, due to the accumulation of one or more negative-scoring residue alignments; or the end of either sequence is reached. The BLAST algorithm parameters W, T, and X determine the sensitivity and speed of the alignment. The BLASTN program (for nucleotide sequences) uses as defaults a wordlength (W) of 11, an expectation (E) of 10, a cutoff of 100, M=5, N=-4, and a comparison of both strands. For amino acid sequences, the BLASTP program uses as defaults a wordlength (W)

of 3, an expectation (E) of 10, and the BLOSUM62 scoring matrix (*see* Henikoff & Henikoff, *Proc. Natl. Acad. Sci. USA* **89**:10915 (1989)).

In addition to calculating percent sequence identity, the BLAST algorithm also performs a statistical analysis of the similarity between two sequences (*see, e.g.*, Karlin & Altschul, *Proc. Nat'l. Acad. Sci. USA* **90**:5873-5787 (1993)). One measure of similarity provided by the BLAST algorithm is the smallest sum probability (P(N)), which provides an indication of the probability by which a match between two nucleotide or amino acid sequences would occur by chance. For example, a nucleic acid is considered similar to a reference sequence if the smallest sum probability in a comparison of the test nucleic acid to the reference nucleic acid is less than about 0.1, more preferably less than about 0.01, and most preferably less than about 0.001.

Another indication that two nucleic acid sequences are substantially identical/homologous is that the two molecules hybridize to each other under stringent conditions. The phrase "hybridizing specifically to," refers to the binding, duplexing, or hybridizing of a molecule only to a particular nucleotide sequence under stringent conditions, including when that sequence is present in a complex mixture (*e.g.*, total cellular) DNA or RNA. "Bind(s) substantially" refers to complementary hybridization between a probe nucleic acid and a target nucleic acid and embraces minor mismatches that can be accommodated by reducing the stringency of the hybridization media to achieve the desired detection of the target polynucleotide sequence.

"Stringent hybridization conditions" and "stringent hybridization wash conditions" in the context of nucleic acid hybridization experiments such as Southern and northern hybridizations are sequence dependent, and are different under different environmental parameters. Longer sequences hybridize specifically at higher temperatures. An extensive guide to the hybridization of nucleic acids is found in Tijssen, LABORATORY TECHNIQUES IN BIOCHEMISTRY AND MOLECULAR BIOLOGY--HYBRIDIZATION WITH NUCLEIC ACID PROBES, part I, chapter 2 "Overview of principles of hybridization and the strategy of nucleic acid probe assays," Elsevier, New York (1993). Generally, highly stringent hybridization and wash conditions are selected to be about 5 °C lower than the thermal melting point (T_m) for the specific sequence at a defined ionic strength and pH. Typically, under "stringent conditions" a probe will hybridize to its target subsequence, but no to unrelated sequences.

The T_m is the temperature (under defined ionic strength and pH) at which 50% of the target sequence hybridizes to a perfectly matched probe. Very stringent conditions are selected to be equal to the T_m for a particular probe. An example of stringent hybridization conditions for hybridization of complementary nucleic acids which have more than 100 complementary residues on a filter in a Southern or northern blot is 50% formamide with 1 mg of heparin at 42 °C, with the hybridization being carried out overnight. An example of highly stringent wash conditions is 0.15M NaCl at 72 °C for about 15 minutes. An example of stringent wash conditions is a 0.2x SSC wash at 65 °C for 15 minutes (*see,* Sambrook, *infra.,* for a description of SSC buffer). Often, a high stringency wash is preceded by a low stringency wash to remove background probe signal. An example medium stringency wash for a duplex of, *e.g.,* more than 100 nucleotides, is 1x SSC at 45°C for 15 minutes. An example low stringency wash for a duplex of, *e.g.,* more than 100 nucleotides, is 4-6x SSC at 40 °C for 15 minutes. For short probes (*e.g.,* about 10 to 50 nucleotides), stringent conditions typically involve salt concentrations of less than about 1.0 M Na ion, typically about 0.01 to 1.0 M Na ion concentration (or other salts) at pH 7.0 to 8.3, and the temperature is typically at least about 30 °C. Stringent conditions can also be achieved with the addition of destabilizing agents such as formamide. In general, a signal to noise ratio of 2x (or higher) than that observed for an unrelated probe in the particular hybridization assay indicates detection of a specific hybridization. Nucleic acids which do not hybridize to each other under stringent conditions are still substantially identical if the polypeptides which they encode are substantially identical. This occurs, *e.g.,* when a copy of a nucleic acid is created using the maximum codon degeneracy permitted by the genetic code.

A further indication that two nucleic acid sequences or polypeptides are substantially identical/homologous is that the polypeptide encoded by the first nucleic acid is immunologically cross reactive with, or specifically binds to, the polypeptide encoded by the second nucleic acid. Thus, a polypeptide is typically substantially identical to a second polypeptide, for example, where the two peptides differ only by conservative substitutions.

"Conservatively modified variations" of a particular polynucleotide sequence refers to those polynucleotides that encode identical or essentially identical amino acid sequences, or where the polynucleotide does not encode an amino acid sequence, to essentially identical sequences. Because of the degeneracy of the genetic code, a large

number of functionally identical nucleic acids encode any given polypeptide. For instance, the codons CGU, CGC, CGA, CGG, AGA, and AGG all encode the amino acid arginine. Thus, at every position where an arginine is specified by a codon, the codon can be altered to any of the corresponding codons described without altering the encoded polypeptide. Such nucleic acid variations are "silent variations," which are one species of "conservatively modified variations." Every polynucleotide sequence described herein which encodes a polypeptide also describes every possible silent variation, except where otherwise noted. One of skill will recognize that each codon in a nucleic acid (except AUG, which is ordinarily the only codon for methionine) can be modified to yield a functionally identical molecule by standard techniques. Accordingly, each "silent variation" of a nucleic acid which encodes a polypeptide is implicit in each described sequence.

Furthermore, one of skill will recognize that individual substitutions, deletions or additions which alter, add or delete a single amino acid or a small percentage of amino acids (typically less than 5%, more typically less than 1%) in an encoded sequence are "conservatively modified variations" where the alterations result in the substitution of an amino acid with a chemically similar amino acid. Conservative substitution tables providing functionally similar amino acids are well known in the art. The following five groups each contain amino acids that are conservative substitutions for one another:

Aliphatic: Glycine (G), Alanine (A), Valine (V), Leucine (L), Isoleucine (I); Aromatic: Phenylalanine (F), Tyrosine (Y), Tryptophan (W); Sulfur-containing: Methionine (M), Cysteine (C); Basic: Arginine (R), Lysine (K), Histidine (H); Acidic: Aspartic acid (D), Glutamic acid (E), Asparagine (N), Glutamine (Q). *See also*, Creighton, PROTEINS, W.H. Freeman and Company (1984). In addition, individual substitutions, deletions or additions which alter, add or delete a single amino acid or a small percentage of amino acids in an encoded sequence are also "conservatively modified variations." Sequences that differ by conservative variations are generally homologous.

A "subsequence" refers to a sequence of nucleic acids or amino acids that comprise a part of a longer sequence of nucleic acids or amino acids (*e.g.*, polypeptide) respectively.

The term "gene" is used broadly to refer to any segment of DNA associated with expression of a given RNA or protein. Thus, genes include regions encoding expressed RNAs (which typically include polypeptide coding sequences) and, often, the regulatory sequences required for their expression. Genes can be obtained from a variety of

sources, including cloning from a source of interest or synthesizing from known or predicted sequence information, and may include sequences designed to have desired parameters.

The term "isolated", when applied to a nucleic acid or protein, denotes that the nucleic acid or protein is essentially free of other cellular components with which it is associated in the natural state.

The term "nucleic acid" refers to deoxyribonucleotides or ribonucleotides and polymers thereof in either single- or double-stranded form. Unless specifically limited, the term encompasses nucleic acids containing known analogues of natural nucleotides which have similar binding properties as the reference nucleic acid and are metabolized in a manner similar to naturally occurring nucleotides. Unless otherwise indicated, a particular nucleic acid sequence also implicitly encompasses conservatively modified variants thereof (*e.g.* degenerate codon substitutions) and complementary sequences and as well as the sequence explicitly indicated. Specifically, degenerate codon substitutions may be achieved by generating sequences in which the third position of one or more selected (or all) codons is substituted with mixed-base and/or deoxyinosine residues (Batzer *et al., Nucleic Acid Res.* **19**: 5081 (1991); Ohtsuka *et al., J. Biol. Chem.* **260**: 2605-2608 (1985); Cassol *et al.* (1992) ; Rossolini *et al., Mol. Cell. Probes* **8**: 91-98 (1994)). The term nucleic acid is generic to the terms "gene", "DNA," "cDNA", "oligonucleotide," "RNA," "mRNA," "polynucleotide" and the like.

"Nucleic acid derived from a gene" refers to a nucleic acid for whose synthesis the gene, or a subsequence thereof, has ultimately served as a template. Thus, an mRNA, a cDNA reverse transcribed from an mRNA, an RNA transcribed from that cDNA, a DNA amplified from the cDNA, an RNA transcribed from the amplified DNA, *etc.*, are all derived from the gene and detection of such derived products is indicative of the presence and/or abundance of the original gene and/or gene transcript in a sample.

A nucleic acid is "operably linked" when it is placed into a functional relationship with another nucleic acid sequence. For instance, a promoter or enhancer is operably linked to a coding sequence if it increases the transcription of the coding sequence.

A "recombinant expression cassette" or simply an "expression cassette" is a nucleic acid construct, generated recombinantly or synthetically, with nucleic acid elements that are capable of effecting expression of a structural gene in hosts compatible with such sequences. Expression cassettes include at least promoters and optionally, transcription

termination signals. Typically, the recombinant expression cassette includes a nucleic acid to be transcribed (*e.g.*, a nucleic acid encoding a desired polypeptide), and a promoter. Additional factors necessary or helpful in effecting expression may also be used as described herein. For example, an expression cassette can also include nucleotide sequences that encode a signal sequence that directs secretion of an expressed protein from the host cell. Transcription termination signals, enhancers, and other nucleic acid sequences that influence gene expression, can also be included in an expression cassette.

"Alkyl" refers to straight- and branched-chain, saturated and unsaturated hydrocarbons. "Lower alkyl", as used herein, refers to "alkyl" groups having from about 1 to about 6 carbon atoms.

"Substituted alkyl" refers to alkyl as just described including one or more functional groups such as lower alkyl, aryl, acyl, halogen (*i.e.*, alkylhalos, *e.g.*, CF_3), hydroxy, amino, alkoxy, alkylamino, acylamino, acyloxy, aryloxy, aryloxyalkyl, mercapto, both saturated and unsaturated cyclic hydrocarbons, heterocycles and the like. These groups may be attached to any carbon of the alkyl moiety.

The term "aryl" is used herein to refer to an aromatic substituent which may be a single aromatic ring or multiple aromatic rings which are fused together, linked covalently, or linked to a common group such as a methylene or ethylene moiety. The common linking group may also be a carbonyl as in benzophenone. The aromatic ring(s) may include phenyl, napthyl, biphenyl, diphenylmethyl and benzophenone among others. The term "aryl" encompasses "arylalkyl."

The term "alkylarene" is used herein to refer to a subset of "aryl" in which the aryl group is substituted with an alkyl group as defined herein.

"Substituted aryl" refers to aryl as just described including one or more functional groups such as lower alkyl, acyl, halogen, alkylhalos (*e.g.* CF_3), hydroxy, amino, alkoxy, alkylamino, acylamino, acyloxy, mercapto and both saturated and unsaturated cyclic hydrocarbons which are fused to the aromatic ring(s), linked covalently or linked to a common group such as a methylene or ethylene moiety. The linking group may also be a carbonyl such as in cyclohexyl phenyl ketone. The term "substituted aryl" encompasses "substituted arylalkyl."

The term "acyl" is used to describe a ketone substituent, —C(O)R, wherein R is alkyl or substituted alkyl, aryl or substituted aryl as defined herein.

The term "halogen" is used herein to refer to fluorine, bromine, chlorine and iodine atoms.

The term "hydroxy" is used herein to refer to the group —OH.

The term "amino" is used to describe primary amines, R—NH_2, wherein R is alkyl or substituted alkyl, aryl or substituted aryl as defined herein.

The term "alkoxy" is used herein to refer to the —OR group, wherein R is a lower alkyl, substituted lower alkyl, aryl, substituted aryl, arylalkyl or substituted arylalkyl wherein the alkyl, aryl, substituted aryl, arylalkyl and substituted arylalkyl groups are as described herein. Suitable alkoxy radicals include, for example, methoxy, ethoxy, phenoxy, substituted phenoxy, benzyloxy, phenethyloxy, t-butoxy, *etc.*

The term "alkylamino" denotes secondary and tertiary amines wherein the alkyl groups may be either the same or different and may consist of straight or branched, saturated or unsaturated hydrocarbons.

The term "unsaturated cyclic hydrocarbon" is used to describe a non-aromatic group with at least one double bond, such as cyclopentene, cyclohexene, *etc.* and substituted analogues thereof.

The term "heteroaryl" as used herein refers to aromatic rings in which one or more carbon atoms of the aromatic ring(s) are substituted by a heteroatom such as nitrogen, oxygen or sulfur. Heteroaryl refers to structures which may be a single aromatic ring, multiple aromatic ring(s), or one or more aromatic rings coupled to one or more non-aromatic ring(s). In structures having multiple rings, the rings can be fused together, linked covalently, or linked to a common group such as a methylene or ethylene moiety. The common linking group may also be a carbonyl as in phenyl pyridyl ketone. As used herein, rings such as thiophene, pyridine, isoxazole, phthalimide, pyrazole, indole, furan, *etc.* or benzo-fused analogues of these rings are defined by the term "heteroaryl."

"Alkylheteroaryl" defines a subset of "heteroaryl" substituted with an alkyl group, as defined herein.

"Substituted heteroaryl" refers to heteroaryl as just described wherein the heteroaryl nucleus is substituted with one or more functional groups such as lower alkyl, acyl, halogen, alkylhalos (*e.g.* CF_3), hydroxy, amino, alkoxy, alkylamino, acylamino, acyloxy, mercapto, *etc.* Thus, substituted analogues of heteroaromatic rings such as thiophene, pyridine, isoxazole, phthalimide, pyrazole, indole, furan, *etc.* or benzo-fused analogues of these rings are defined by the term "substituted heteroaryl."

The term "heterocyclic" is used herein to describe a saturated or unsaturated non-aromatic group having a single ring or multiple condensed rings from about 1 to about 12 carbon atoms and from about 1 to about 4 heteroatoms selected from nitrogen, sulfur or oxygen within the ring. Such heterocycles are, for example, tetrahydrofuran, morpholine, piperidine, pyrrolidine, *etc.*

The term "substituted heterocyclic" as used herein describes a subset of "heterocyclic" wherein the heterocycle nucleus is substituted with one or more functional groups such as lower alkyl, acyl, halogen, alkylhalos (*e.g.*, CF_3), hydroxy, amino, alkoxy, alkylamino, acylamino, acyloxy, mercapto, *etc.*

The term "alkylheterocyclic" defines a subset of "heterocyclic" substituted with an alkyl group, as defined herein.

The term "substituted heterocyclicalkyl" defines a subset of "heterocyclic alkyl" wherein the heterocyclic nucleus is substituted with one or more functional groups such as lower alkyl, acyl, halogen, alkylhalos (*e.g.* CF_3), hydroxy, amino, alkoxy, alkylamino, acylamino, acyloxy, mercapto, *etc.*

Introduction

This invention describes the generation of evolved dioxygenases with enhanced performance for use in the production of chemicals of industrial interest using any of a variety of shuffling techniques, including, for example, gene, family and whole genome shuffling as described herein. In this invention, shuffling is used to enhance properties of dioxygenases, such as forward rate kinetics, substrate specificity and affinity and also to decrease susceptibility of dioxygenases to reversible inhibitors and inactivation by solvents, starting materials and reaction products and intermediates generated during the catalytic cycle.

While much of the discussion below deals explicitly with arene dioxygenases, this is for clarity of illustration. The discussion is representative of the chemistries and improvements which can be made to other useful dioxygenases, such as the structurally and functionally similar peroxidases and chlorperoxidases, as well as to the structurally unrelated iron-sulfur methane dioxygenases and other enzymes noted herein using shuffling methodologies, e.g., shuffling of a single gene or a gene family, as described herein.

In a first aspect, the present invention provides a method for obtaining a nucleic acid that encodes an improved polypeptide possessing dioxygenase activity. The improved polypeptide has at least one property improved over a naturally occurring dioxygenase polypeptide. The method includes: (a) creating a library of recombinant polynucleotides encoding a recombinant dioxygenase polypeptide; and (b) screening the library to identify a recombinant polynucleotide that encodes an improved recombinant dioxygenase polypeptide that has at least one property improved over the naturally occurring polypeptide. Also provided are nucleic acids produced by this method that encode a dioxygenase polypeptide having at least one property improved over a naturally occurring dioxygenase polypeptide.

In a preferred embodiment, the nucleic acid libraries of the invention are constructed by a method that includes shuffling a plurality of parental polynucleotides to produce one or more recombinant dioxygenase polynucleotide encoding the improved property. In another preferred embodiment, the polynucleotides are homologous. A detailed description of shuffling techniques is provided in Part A, herein below.

In another embodiment, at least one of the parental polynucleotides is selected from polynucleotides that encode at least one dioxygenase activity and those that do not encode at least one dioxygenase activity. Typically, the parental dioxygenase polynucleotide encodes a complete polypeptide or a polypeptide fragment selected from an arene dioxygenase or fragments thereof.

In a preferred embodiment, the activity catalyzed by a dioxygenase polypeptide is a member selected from one or more reactions described in Figures 14A-14R. Other oxidative transformations will be apparent to those of skill in the art.

The invention provides significant advantages over previously used methods for optimization of dioxygenase genes. For example, nucleic acid shuffling can result in optimization of a desirable property even in the absence of a detailed understanding of the mechanism by which the particular property is mediated. In addition, entirely new properties can be obtained upon shuffling of DNAs, *i.e.*, shuffled DNAs can encode polypeptides or RNAs with properties entirely absent in the parental DNAs which are shuffled.

The properties or characteristics that can be acquired or improved vary widely, and depend on the choice of substrate. For example, for dioxygenase genes, properties that one can improve include, but are not limited to, increased range of

dioxygenase activity encoded by a particular gene, increased potency against a dioxygenase target, increased expression level of the dioxygenase gene, increased tolerance of the protein encoded by the dioxygenase gene to protease degradation (or other natural protein or RNA degradative processes), increased dioxygenase activity ranges for conditions such as heat, cold, low or high pH, reduced toxicity to the host cell, and increased resistance of the polypeptide and/or the organism expressing the polypeptide to organic solvents, and reaction feedstocks, intermediates and products.

The targets for modification vary in different applications, as does the property sought to be acquired or improved. Examples of candidate targets for acquisition of a property or improvement in a property include genes that encode proteins which have enzymatic or other activities useful in dioxygenase reactions.

The methods typically use at least two variant forms of a starting target. The variant forms of candidate substrates can show substantial sequence or secondary structural similarity with each other, but they should also differ in at least one and preferably at least two positions.

The initial diversity between forms can be the result of natural variation, *e.g.*, the different variant forms (homologs) are obtained from different individuals or strains of an organism, or constitute related sequences from the same organism (*e.g.*, allelic variations), or constitute homologs from different organisms (interspecific variants). Alternatively, initial diversity can be induced, *e.g.*, the variant forms can be generated by error-prone transcription, such as an error-prone PCR or use of a polymerase which lacks proof-reading activity (*see,* Liao (1990) *Gene* 88:107-111), of the first variant form, or, by replication of the first form in a mutator strain (mutator host cells are discussed in further detail below, and are generally well known). The initial diversity between substrates is greatly augmented in subsequent steps of recombination for library generation.

A mutator strain can include any mutants in any organism impaired in the functions of mismatch repair. These include mutant gene products of mutS, mutT, mutH, mutL, ovrD, dcm, vsr, umuC, umuD, sbcB, recJ, *etc.* The impairment is achieved by genetic mutation, allelic replacement, selective inhibition by an added reagent such as a small molecule or an expressed antisense RNA, or other techniques. Impairment can be of the genes noted, or of homologous genes in any organism.

Therefore, in carrying out the practice of the present invention, at least two variant forms of a nucleic acid which can confer dioxygenase activity are recombined to

produce a library of recombinant dioxygenase genes. The library is then screened to identify at least one recombinant dioxygenase gene that is optimized for the particular property or properties of interest.

The parental polynucleotides can be shuffled in substantially any cell type, including prokaryotes, eukaryotes, yeast, bacteria and fungi. In a preferred embodiment, the one or more recombinant dioxygenase nucleic acid is present in one or more bacterial, yeast, or fungal cells and the method involves: pooling multiple separate dioxygenase nucleic acids; screening the resulting pooled dioxygenase nucleic acids to identify a distinct or improved recombinant dioxygenase nucleic acid that exhibits distinct or improved dioxygenase activity compared to a non-recombinant dioxygenase activity nucleic acid; and cloning the distinct or improved recombinant nucleic acid.

Often, improvements are achieved after one round of recombination and selection. However, recursive sequence recombination can be employed to achieve still further improvements in a desired property, or to bring about new (or "distinct") properties. Recursive sequence recombination entails successive cycles of recombination to generate molecular diversity. That is, one creates a family of nucleic acid molecules showing some sequence identity to each other but differing in the presence of mutations. In any given cycle, recombination can occur *in vivo* or *in vitro*, intracellularly or extracellularly. Furthermore, diversity resulting from recombination can be augmented in any cycle by applying prior methods of mutagenesis (*e.g.*, error-prone PCR or cassette mutagenesis) to either the substrates or products for recombination.

A recombination cycle is usually followed by at least one cycle of screening or selection for molecules having a desired property or characteristic. If a recombination cycle is performed *in vitro*, the products of recombination, *i.e.*, recombinant segments, are sometimes introduced into cells before the screening step. Recombinant segments can also be linked to an appropriate vector or other regulatory sequences before screening. Alternatively, products of recombination generated *in vitro* are sometimes packaged in viruses (*e.g.*, bacteriophage) before screening. If recombination is performed *in vivo*, recombination products can sometimes be screened in the cells in which recombination occurred. In other applications, recombinant segments are extracted from the cells, and optionally packaged as viruses, before screening.

The nature of screening or selection depends on what property or characteristic is to be acquired or the property or characteristic for which improvement is

sought, and many examples are discussed below. It is not usually necessary to understand the molecular basis by which particular products of recombination (recombinant segments) have acquired new or improved properties or characteristics relative to the starting substrates. For example, a dioxygenase gene can have many component sequences each having a different intended role (*e.g.*, coding sequence, regulatory sequences, targeting sequences, stability-conferring sequences, subunit sequences and sequences affecting integration). Each of these component sequences can be varied and recombined simultaneously. Screening/selection can then be performed, for example, for recombinant segments that have increased ability to confer dioxygenase activity upon a cell without the need to attribute such improvement to any of the individual component sequences of the vector.

Depending on the particular screening protocol used for a desired property, initial round(s) of screening are sometimes performed using bacterial cells due to high transfection efficiencies and ease of culture. However, for eukaryotic dioxygenases such as eukaryotic arene dioxygenases, bacterial expression is often not practical, and yeast, fungal or other eukaryotic systems are used for library expression and screening. Similarly, other types of screening that are not amenable to screening in bacterial or simple eukaryotic library cells, are performed in cells selected for use in an environment close to that of their intended use. Final rounds of screening can be performed in the precise cell type of intended use.

If further improvement in a property is desired, at least one and usually a collection of recombinant segments surviving a first round of screening/selection are subject to a further round of recombination. These recombinant segments can be recombined with each other or with exogenous segments representing the original substrates or further variants thereof. Again, recombination can proceed *in vitro* or *in vivo*. If the previous screening step identifies desired recombinant segments as components of cells, the components can be subjected to further recombination *in vivo*, or can be subjected to further recombination *in vitro*, or can be isolated before performing a round of *in vitro* recombination. Conversely, if the previous screening step identifies desired recombinant segments in naked form or as components of viruses, these segments can be introduced into cells to perform a round of *in vivo* recombination. The second round of recombination, irrespective how performed, generates further recombinant segments which encompass additional diversity than is present in recombinant segments resulting from previous rounds.

The second round of recombination is optionally followed by a further round of screening/selection according to the principles discussed above for the first round. The stringency of screening/selection can be increased between rounds. Also, the nature of the screen and the property being screened for can vary between rounds if improvement in more than one property is desired or if acquiring more than one new property is desired. Additional rounds of recombination and screening can then be performed until the recombinant segments have sufficiently evolved to acquire the desired new or improved property or function.

In a preferred embodiment, the invention provides a recursive method for making a nucleic acid encoding a specific dioxygenase activity. In this method, the parental nucleic acids are shuffled in a plurality of cells and the method optionally further includes one or more of: (a) recombining DNA from the plurality of cells that display dioxygenase activity with a library of DNA fragments, at least one of which undergoes recombination with a segment in a cellular DNA present in the cells to produce recombined cells, or recombining DNA between the plurality of cells that display dioxygenase activity to produce cells with modified dioxygenase activity; (b) recombining and screening the recombined or modified cells to produce further recombined cells that have evolved additionally modified dioxygenase activity; and, (c) repeating (a) or (b) until the further recombined cells have acquired a desired dioxygenase activity.

In another preferred embodiment, the invention provides a method for making a nucleic acid encoding a specific dioxygenase activity. This method includes: (a) recombining at least one distinct or improved recombinant nucleic acid with a further dioxygenase activity nucleic acid, which further nucleic acid is the same or different from one or more of the plurality of parental nucleic acids to produce a library of recombinant dioxygenase nucleic acids; (b) screening the library to identify at least one further distinct or improved recombinant dioxygenase nucleic acid that exhibits a further improvement or distinct property compared to the plurality of parental nucleic acids; and, optionally; (c) repeating (a) and (b) until the resulting further distinct or improved recombinant nucleic acid shows an additionally distinct or improved dioxygenase property.

The practice of this invention involves the construction of recombinant nucleic acids and the expression of genes in transfected host cells. Molecular cloning techniques to achieve these ends are known in the art. A wide variety of cloning and *in vitro* amplification methods suitable for the construction of recombinant nucleic acids such

as expression vectors are well-known to persons of skill. General texts which describe molecular biological techniques useful herein, including mutagenesis, include Berger and Kimmel, GUIDE TO MOLECULAR CLONING TECHNIQUES, METHODS IN ENZYMOLOGY, volume 152 Academic Press, Inc., San Diego, CA (Berger); Sambrook *et al.*, MOLECULAR CLONING - A LABORATORY MANUAL (2nd Ed.), Vol. 1-3, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, 1989 ("Sambrook") and CURRENT PROTOCOLS IN MOLECULAR BIOLOGY, F.M. Ausubel *et al.*, eds., Current Protocols, a joint venture between Greene Publishing Associates, Inc. and John Wiley & Sons, Inc., (supplemented through 1998) ("Ausubel")). Examples of techniques sufficient to direct persons of skill through *in vitro* amplification methods, including the polymerase chain reaction (PCR) the ligase chain reaction (LCR), Qβ-replicase amplification and other RNA polymerase mediated techniques (*e.g.*, NASBA) are found in Berger, Sambrook, and Ausubel, as well as Mullis *et al.*, (1987) U.S. Patent No. 4,683,202; PCR PROTOCOLS A GUIDE TO METHODS AND APPLICATIONS (Innis *et al.* eds) Academic Press Inc., San Diego, CA (1990) (Innis); Arnheim & Levinson (October 1, 1990) *C&EN* 36-47; *The Journal Of NIH Research* **3**:81-94 (1991); (Kwoh *et al.*, *Proc. Natl. Acad. Sci. USA* **86**:1173; (1989) Guatelli *et al.*, *Proc. Natl. Acad. Sci. USA* **87**:1874 (1990); Lomell *et al.*, *J. Clin. Chem* **35**:1826 (1989); Landegren *et al.*, *Science* **241**:1077-1080 (1988); Van Brunt, *Biotechnology* **8**:291-294 (1990); Wu and Wallace, *Gene* **4**:560 (1989); Barringer *et al.*, *Gene* **89**:117 (1990), and Sooknanan and Malek, *Biotechnology* **13**:563-564 (1995). Improved methods of cloning *in vitro* amplified nucleic acids are described in Wallace *et al.*, U.S. Pat. No. 5,426,039. Improved methods of amplifying large nucleic acids by PCR are summarized in Cheng *et al.*, *Nature* **369**:684-685 (1994), and the references cited therein, in which PCR amplicons of up to 40kb are generated. One of skill will appreciate that essentially any RNA can be converted into a double stranded DNA suitable for restriction digestion, PCR expansion and sequencing using reverse transcriptase and a polymerase. *See*, Ausbel, Sambrook and Berger, *all supra.*

In another aspect, the present invention provides a method of increasing dioxygenase activity in a cell. The method includes performing whole genome shuffling of a plurality of genomic nucleic acids in the cell and selecting for one or more dioxygenase activity. In this aspect of the invention, the genomic nucleic acids can be from substantially any source. In a preferred embodiment of this aspect of the invention, the genomic nucleic

acids are from a species or strain different from the cell. In a further preferred embodiment, the cell is of prokaryotic or eukaryotic origin.

Substantially any dioxygenase property can be selected for using the methods of the invention. A preferred property is the activity of the polypeptide towards a particular class of substrates. In a preferred embodiment, the dioxygenase property is its ability to effect one or more reactions described in Figures 14A-14R.

In a third aspect, the invention provides a nucleic acid shuffling mixture comprising: at least three homologous DNAs, each of which is derived from a nucleic acid encoding a polypeptide or polypeptide fragment which encodes dioxygenase activity. In a preferred embodiment of this aspect of the invention, the at least three homologous DNAs are present in cell culture or *in vitro*.

Oligonucleotides for use as probes, *e.g.*, in *in vitro* amplification methods, for use as gene probes, or as shuffling targets (*e.g.*, synthetic genes or gene segments) are typically synthesized chemically according to the solid phase phosphoramidite triester method described by Beaucage and Caruthers, *Tetrahedron Letts.*, **22(20)**:1859-1862 (1981), *e.g.*, using an automated synthesizer, as described in Needham-VanDevanter *et al.*, *Nucleic Acids Res.* **12**:6159-6168 (1984). Oligonucleotides can also be custom made and ordered from a variety of commercial sources known to persons of skill.

A. Formats for Sequence Recombination

The methods of the invention entail performing recombination ("shuffling") and screening or selection to "evolve" individual genes, whole plasmids or viruses, multigene clusters, or even whole genomes (Stemmer, *Bio/Technology* **13**:549-553 (1995)). Reiterative cycles of recombination and screening/selection can be performed to further evolve the nucleic acids of interest. Such techniques do not require the extensive analysis and computation required by conventional methods for polypeptide engineering. Shuffling allows the recombination of large numbers of mutations in a minimum number of selection cycles, in contrast to natural pair-wise recombination events (*e.g.*, as occur during sexual replication). Thus, the sequence recombination techniques described herein provide particular advantages in that they provide recombination between mutations in any or all of these, thereby providing a very fast way of exploring the manner in which different combinations of mutations can affect a desired result. In some instances, however, structural and/or functional information is available which, although not required for sequence recombination, provides opportunities for modification of the technique.

A variety of nucleic acid shuffling protocols are available and fully described in the art. Descriptions of a variety of shuffling methods for generating modified nucleic acid sequences for use in the methods of the present invention include the following publications and the references cited therein: Stemmer, et al. (1999) "Molecular breeding of viruses for targeting and other clinical properties" Tumor Targeting 4:1-4; Ness et al. (1999) "DNA Shuffling of subgenomic sequences of subtilisin" Nature Biotechnology 17:893-896; Chang et al. (1999) "Evolution of a cytokine using DNA family shuffling" Nature Biotechnology 17:793-797; Minshull and Stemmer (1999) "Protein evolution by molecular breeding" Current Opinion in Chemical Biology 3:284-290; Christians et al. (1999) "Directed evolution of thymidine kinase for AZT phosphorylation using DNA family shuffling" Nature Biotechnology 17:259-264; Crameri et al. (1998) "DNA shuffling of a family of genes from diverse species accelerates directed evolution" Nature 391:288-291; Crameri et al. (1997) "Molecular evolution of an arsenate detoxification pathway by DNA shuffling," Nature Biotechnology 15:436-438; Zhang et al. (1997) "Directed evolution of an effective fucosidase from a galactosidase by DNA shuffling and screening" Proc. Natl. Acad. Sci. USA 94:4504-4509; Patten et al. (1997) "Applications of DNA Shuffling to Pharmaceuticals and Vaccines" Current Opinion in Biotechnology 8:724-733; Crameri et al. (1996) "Construction and evolution of antibody-phage libraries by DNA shuffling" Nature Medicine 2:100-103; Crameri et al. (1996) "Improved green fluorescent protein by molecular evolution using DNA shuffling" Nature Biotechnology 14:315-319; Gates et al. (1996) "Affinity selective isolation of ligands from peptide libraries through display on a lac repressor 'headpiece dimer'" Journal of Molecular Biology 255:373-386; Stemmer (1996) "Sexual PCR and Assembly PCR" In: The Encyclopedia of Molecular Biology. VCH Publishers, New York. pp.447-457; Crameri and Stemmer (1995) "Combinatorial multiple cassette mutagenesis creates all the permutations of mutant and wildtype cassettes" BioTechniques 18:194-195; Stemmer et al., (1995) "Single-step assembly of a gene and entire plasmid form large numbers of oligodeoxy-ribonucleotides" Gene, 164:49-53; Stemmer (1995) "The Evolution of Molecular Computation" Science 270: 1510; Stemmer (1995) "Searching Sequence Space" Bio/Technology 13:549-553; Stemmer (1994) "Rapid evolution of a protein in vitro by DNA shuffling" Nature 370:389-391; and Stemmer (1994) "DNA shuffling by random fragmentation and reassembly: In vitro recombination for molecular evolution." Proc. Natl. Acad. Sci. USA 91:10747-10751.

Additional details regarding DNA shuffling methods can be found in the following U.S. patents, PCT publications, and EPO publications: USPN 5,605,793 to Stemmer (February 25, 1997), "Methods for In Vitro Recombination;" USPN 5,811,238 to Stemmer et al. (September 22, 1998) "Methods for Generating Polynucleotides having Desired Characteristics by Iterative Selection and Recombination;" USPN 5,830,721 to Stemmer et al. (November 3, 1998), "DNA Mutagenesis by Random Fragmentation and Reassembly;" USPN 5,834,252 to Stemmer, et al. (November 10, 1998) "End-Complementary Polymerase Reaction;" USPN 5,837,458 to Minshull, et al. (November 17, 1998), "Methods and Compositions for Cellular and Metabolic Engineering;" WO 95/22625, Stemmer and Crameri, "Mutagenesis by Random Fragmentation and Reassembly;" WO 96/33207 by Stemmer and Lipschutz "End Complementary Polymerase Chain Reaction;" WO 97/20078 by Stemmer and Crameri "Methods for Generating Polynucleotides having Desired Characteristics by Iterative Selection and Recombination;" WO 97/35966 by Minshull and Stemmer, "Methods and Compositions for Cellular and Metabolic Engineering;" WO 99/41402 by Punnonen et al. "Targeting of Genetic Vaccine Vectors;" WO 99/41383 by Punnonen et al. "Antigen Library Immunization;" WO 99/41369 by Punnonen et al. "Genetic Vaccine Vector Engineering;" WO 99/41368 by Punnonen et al. "Optimization of Immunomodulatory Properties of Genetic Vaccines;" EP 752008 by Stemmer and Crameri, "DNA Mutagenesis by Random Fragmentation and Reassembly;" EP 0932670 by Stemmer "Evolving Cellular DNA Uptake by Recursive Sequence Recombination;" WO 99/23107 by Stemmer et al., "Modification of Virus Tropism and Host Range by Viral Genome Shuffling;" WO 99/21979 by Apt et al., "Human Papillomavirus Vectors;" WO 98/31837 by del Cardayre et al. "Evolution of Whole Cells and Organisms by Recursive Sequence Recombination;" WO 98/27230 by Patten and Stemmer, "Methods and Compositions for Polypeptide Engineering;" WO 98/13487 by Stemmer et al., "Methods for Optimization of Gene Therapy by Recursive Sequence Shuffling and Selection," WO 00/00632, "Methods for Generating Highly Diverse Libraries," WO 00/09679, "Methods for Obtaining in Vitro Recombined Polynucleotide Sequence Banks and Resulting Sequences," WO 98/42832 by Arnold et al., "Recombination of Polynucleotide Sequences Using Random or Defined Primers," WO 99/29902 by Arnold et al., "Method for Creating Polynucleotide and Polypeptide Sequences," WO 98/41653 by Vind, "An in Vitro Method for Construction of a DNA

Library," and WO 98/41622 by Borchert et al., "Method for Constructing a Library Using DNA Shuffling."

Certain U.S. applications provide additional details regarding shuffling methods, including "SHUFFLING OF CODON ALTERED GENES" by Patten et al. filed September 29, 1998, (USSN 60/102,362), January 29, 1999 (USSN 60/117,729), and September 28, 1999, (USSN 09/407,800); "EVOLUTION OF WHOLE CELLS AND ORGANISMS BY RECURSIVE SEQUENCE RECOMBINATION", by del Cardayre et al. filed July 15, 1998 (USSN 09/166,188), and July 15, 1999 (USSN 09/354,922); "OLIGONUCLEOTIDE MEDIATED NUCLEIC ACID RECOMBINATION" by Crameri et al., filed February 5, 1999 (USSN 60/118,813), June 24, 1999 (USSN 60/141,049), and September 28, 1999 (USSN 09/408,392); "USE OF CODON-BASED OLIGONUCLEOTIDE SYNTHESIS FOR SYNTHETIC SHUFFLING" by Welch et al., filed September 28, 1999 (USSN 09/408,393); "METHODS FOR MAKING CHARACTER STRINGS, POLYNUCLEOTIDES & POLYPEPTIDES HAVING DESIRED CHARACTERISTICS" by Selifonov and Stemmer, filed February 5, 1999 (USSN 60/118854) and October 12, 1999 (USSN 09/416,375); and "SINGLE-STRANDED NUCLEIC ACID TEMPLATE-MEDIATED RECOMBINATION AND NUCLEIC ACID FRAGMENT ISOLATION" by Affholter, USSN 60/186,482 filed March 2,2000.

In brief, a variety of shuffling formats are applicable to the present invention and set forth, e.g., in the references above. The following exemplify some of the different types of formats. First, nucleic acids can be recombined in vitro by any of a variety of techniques discussed in the references above, including e.g., DNAse digestion of nucleic acids to be recombined followed by ligation and/or PCR reassembly of the nucleic acids. Second, nucleic acids can be recursively recombined in vivo, e.g., by allowing recombination to occur between nucleic acids in cells. Third, whole genome recombination methods can be used in which whole genomes of cells or other organisms are recombined, optionally including spiking of the genomic recombination mixtures with desired library components (e.g., genes corresponding to the pathways of the present invention). Fourth, synthetic recombination methods can be used, in which oligonucleotides corresponding to targets of interest are synthesized and reassembled in PCR or ligation reactions which include oligonucleotides which correspond to more than one parental nucleic acid, thereby generating new recombined nucleic acids. Oligonucleotides can be made by standard nucleotide addition methods, or can be made, e.g., by tri-nucleotide synthetic approaches.

Fifth, in silico methods of recombination can be effected in which genetic algorithms are used in a computer to recombine sequence strings which correspond to homologous (or even non-homologous) nucleic acids. The resulting recombined sequence strings are optionally converted into nucleic acids by synthesis of nucleic acids which correspond to the recombined sequences, e.g., in concert with oligonucleotide synthesis/ gene reassembly techniques. Any of the preceding general recombination formats can be practiced in a reiterative fashion to generate a more diverse set of recombinant nucleic acids. Sixth, methods of accessing natural diversity, e.g., by hybridization of diverse nucleic acids or nucleic acid fragments to single-stranded templates, followed by polymerization and/or ligation to regenerate full-length sequences, optionally followed by degradation of the templates and recovery of the resulting modified nucleic acids can be used.

The above references provide these and other basic recombination formats as well as many modifications of these formats. Regardless of the shuffling format which is used, the nucleic acids of the invention can be recombined (with each other, or with related (or even unrelated) sequences) to produce a diverse set of recombinant nucleic acids, including, e.g., sets of homologous nucleic acids.

Following recombination, any nucleic acids which are produced can be selected for a desired activity. In the context of the present invention, this can include testing for and identifying any activity that can be detected e.g., in an automatable format, by any of the assays in the art. A variety of related (or even unrelated) properties can be assayed for, using any available assay.

DNA mutagenesis and shuffling provide a robust, widely applicable, means of generating diversity useful for the engineering of proteins, pathways, cells and organisms with improved characteristics. In addition to the basic formats described above, it is sometimes desirable to combine shuffling methodologies with other techniques for generating diversity. In conjunction with (or separately from) shuffling methods, a variety of diversity generation methods can be practiced and the results (i.e., diverse populations of nucleic acids) screened for in the systems of the invention. Additional diversity can be introduced by methods which result in the alteration of individual nucleotides or groups of contiguous or non-contiguous nucleotides, i.e., mutagenesis methods. Many mutagenesis methods are found in the above-cited references; additional details regarding mutagenesis methods can be found in the references listed below.

Mutagenesis methods include, for example, those described in PCT/US98/05223; Publ. No. WO98/42727; site-directed mutagenesis (Ling et al. (1997) "Approaches to DNA mutagenesis: an overview" Anal Biochem. 254(2): 157-178; Dale et al. (1996) "Oligonucleotide-directed random mutagenesis using the phosphorothioate method" Methods Mol. Biol. 57:369-374; Smith (1985) "In vitro mutagenesis" Ann. Rev. Genet. 19:423-462; Botstein & Shortle (1985) "Strategies and applications of in vitro mutagenesis" Science 229:1193-1201; Carter (1986) "Site-directed mutagenesis" Biochem. J. 237:1-7; and Kunkel (1987) "The efficiency of oligonucleotide directed mutagenesis" in Nucleic Acids & Molecular Biology (Eckstein, F. and Lilley, D.M.J. eds., Springer Verlag, Berlin)); mutagenesis using uracil containing templates (Kunkel (1985) "Rapid and efficient site-specific mutagenesis without phenotypic selection" Proc. Natl. Acad. Sci. USA 82:488-492; Kunkel et al. (1987) "Rapid and efficient site-specific mutagenesis without phenotypic selection" Methods in Enzymol. 154, 367-382; and Bass et al. (1988) "Mutant Trp repressors with new DNA-binding specificities" Science 242:240-245); oligonucleotide-directed mutagenesis (Methods in Enzymol. 100: 468-500 (1983); Methods in Enzymol. 154: 329-350 (1987); Zoller & Smith (1982) "Oligonucleotide-directed mutagenesis using M13-derived vectors: an efficient and general procedure for the production of point mutations in any DNA fragment" Nucleic Acids Res. 10:6487-6500; Zoller & Smith (1983) "Oligonucleotide-directed mutagenesis of DNA fragments cloned into M13 vectors" Methods in Enzymol. 100:468-500; and Zoller & Smith (1987) "Oligonucleotide-directed mutagenesis: a simple method using two oligonucleotide primers and a single-stranded DNA template" Methods in Enzymol. 154:329-350); phosphorothioate-modified DNA mutagenesis (Taylor et al. (1985) "The use of phosphorothioate-modified DNA in restriction enzyme reactions to prepare nicked DNA" Nucl. Acids Res. 13: 8749-8764; Taylor et al. (1985) "The rapid generation of oligonucleotide-directed mutations at high frequency using phosphorothioate-modified DNA" Nucl. Acids Res. 13: 8765-8787 (1985); Nakamaye & Eckstein (1986) "Inhibition of restriction endonuclease Nci I cleavage by phosphorothioate groups and its application to oligonucleotide-directed mutagenesis" Nucl. Acids Res. 14: 9679-9698; Sayers et al. (1988) "Y-T Exonucleases in phosphorothioate-based oligonucleotide-directed mutagenesis" Nucl. Acids Res. 16:791-802; and Sayers et al. (1988) "Strand specific cleavage of phosphorothioate-containing DNA by reaction with restriction endonucleases in the presence of ethidium bromide" Nucl. Acids Res. 16: 803-814); mutagenesis using gapped duplex DNA (Kramer et al. (1984) "The gapped duplex

DNA approach to oligonucleotide-directed mutation construction" Nucl. Acids Res. 12: 9441-9456; Kramer & Fritz (1987) Methods in Enzymol. "Oligonucleotide-directed construction of mutations via gapped duplex DNA" 154:350-367; Kramer et al. (1988) "Improved enzymatic in vitro reactions in the gapped duplex DNA approach to oligonucleotide-directed construction of mutations" Nucl. Acids Res. 16: 7207; and Fritz et al. (1988) "Oligonucleotide-directed construction of mutations: a gapped duplex DNA procedure without enzymatic reactions in vitro" Nucl. Acids Res. 16: 6987-6999).

Additional suitable methods include point mismatch repair (Kramer et al. (1984) "Point Mismatch Repair" Cell 38:879-887), mutagenesis using repair-deficient host strains (Carter et al. (1985) "Improved oligonucleotide site-directed mutagenesis using M13 vectors" Nucl. Acids Res. 13: 4431-4443; and Carter (1987) "Improved oligonucleotide-directed mutagenesis using M13 vectors" Methods in Enzymol. 154: 382-403), deletion mutagenesis (Eghtedarzadeh & Henikoff (1986) "Use of oligonucleotides to generate large deletions" Nucl. Acids Res. 14: 5115), restriction-selection and restriction-selection and restriction-purification (Wells et al. (1986) "Importance of hydrogen-bond formation in stabilizing the transition state of subtilisin" Phil. Trans. R. Soc. Lond. A 317: 415-423), mutagenesis by total gene synthesis (Nambiar et al. (1984) "Total synthesis and cloning of a gene coding for the ribonuclease S protein" Science 223: 1299-1301; Sakamar and Khorana (1988) "Total synthesis and expression of a gene for the a-subunit of bovine rod outer segment guanine nucleotide-binding protein (transducin)" Nucl. Acids Res. 14: 6361-6372; Wells et al. (1985) "Cassette mutagenesis: an efficient method for generation of multiple mutations at defined sites" Gene 34:315-323; and Grundström et al. (1985) "Oligonucleotide-directed mutagenesis by microscale 'shot-gun' gene synthesis" Nucl. Acids Res. 13: 3305-3316), double-strand break repair (Mandecki (1986) "Oligonucleotide-directed double-strand break repair in plasmids of *Escherichia coli*: a method for site-specific mutagenesis" Proc. Natl. Acad. Sci. USA, 83:7177-7181). Additional details on many of the above methods can be found in Methods in Enzymology Volume 154, which also describes useful controls for trouble-shooting problems with various mutagenesis methods.

In one aspect of the present invention, error-prone PCR can be used to generate nucleic acid variants. Using this technique, PCR is performed under conditions where the copying fidelity of the DNA polymerase is low, such that a high rate of point mutations is obtained along the entire length of the PCR product. Examples of such

techniques are found in the references above and, e.g., in Leung et al. (1989) Technique 1:11-15 and Caldwell et al. (1992) PCR Methods Applic. 2:28-33. Similarly, assembly PCR can be used, in a process which involves the assembly of a PCR product from a mixture of small DNA fragments. A large number of different PCR reactions can occur in parallel in the same vial, with the products of one reaction priming the products of another reaction. Sexual PCR mutagenesis can be used in which homologous recombination occurs between DNA molecules of different but related DNA sequence in vitro, by random fragmentation of the DNA molecule based on sequence homology, followed by fixation of the crossover by primer extension in a PCR reaction. This process is described in the references above, e.g., in Stemmer (1994) Proc. Natl. Acad. Sci. USA 91:10747-10751. Recursive ensemble mutagenesis can be used in which an algorithm for protein mutagenesis is used to produce diverse populations of phenotypically related mutants whose members differ in amino acid sequence. This method uses a feedback mechanism to control successive rounds of combinatorial cassette mutagenesis. Examples of this approach are found in Arkin & Youvan (1992) Proc. Natl. Acad. Sci. USA 89:7811-7815.

As noted, oligonucleotide directed mutagenesis can be used in a process which allows for the generation of site-specific mutations in any nucleic acid sequence of interest. Examples of such techniques are found in the references above and, e.g., in Reidhaar-Olson et al. (1988) Science, 241:53-57. Similarly, cassette mutagenesis can be used in a process which replaces a small region of a double stranded DNA molecule with a synthetic oligonucleotide cassette that differs from the native sequence. The oligonucleotide can contain, e.g., completely and/or partially randomized native sequence(s).

In vivo mutagenesis can be used in a process of generating random mutations in any cloned DNA of interest which involves the propagation of the DNA, e.g., in a strain of *E. coli* that carries mutations in one or more of the DNA repair pathways. These "mutator" strains have a higher random mutation rate than that of a wild-type parent. Propagating the DNA in one of these strains will eventually generate random mutations within the DNA.

Exponential ensemble mutagenesis can be used for generating combinatorial libraries with a high percentage of unique and functional mutants, where small groups of residues are randomized in parallel to identify, at each altered position, amino acids which lead to functional proteins. Examples of such procedures are found in Delegrave & Youvan

(1993) Biotechnology Research 11:1548-1552. Similarly, random and site-directed mutagenesis can be used. Examples of such procedures are found in Arnold (1993) Current Opinion in Biotechnology 4:450-455.

Kits for mutagenesis are also commercially available. For example, kits are available from, e.g., Stratagene (e.g., QuickChange™ site-directed mutagenesis kit; and Chameleon™ double-stranded, site-directed mutagenesis kit), Bio/Can Scientific, Bio-Rad (e.g., using the Kunkel method described above), Boehringer Mannheim Corp., Clonetech Laboratories, DNA Technologies, Epicentre Technologies (e.g., 5 prime 3 prime kit); Genpak Inc, Lemargo Inc, Life Technologies (Gibco BRL), New England Biolabs, Pharmacia Biotech, Promega Corp., Quantum Biotechnologies, Amersham International plc (e.g., using the Eckstein method above), and Anglian Biotechnology Ltd (e.g., using the Carter/Winter method above).

Any of the described shuffling or mutagenesis techniques can be used in conjunction with procedures which introduce additional diversity into a genome, e.g. a bacterial, fungal, animal or plant genome. For example, in addition to the methods above, techniques have been proposed which produce nucleic acid multimers suitable for transformation into a variety of species (*see,* e.g., Schellenberger U.S. Patent No. 5,756,316 and the references above). When such multimers consist of genes that are divergent with respect to one another, (e.g., derived from natural diversity or through application of site directed mutagenesis, error prone PCR, passage through mutagenic bacterial strains, and the like), are transformed into a suitable host, this provides a source of nucleic acid diversity for DNA diversification.

Multimers transformed into host species are suitable as substrates for in vivo shuffling protocols. Alternatively, a multiplicity of polynucleotides sharing regions of partial sequence similarity can be transformed into a host species and recombined in vivo by the host cell. Subsequent rounds of cell division can be used to generate libraries, members of which, comprise a single, homogenous population of monomeric or pooled nucleic acid. Alternatively, the monomeric nucleic acid can be recovered by standard techniques and recombined in any of the described shuffling formats.

Shuffling formats employing chain termination methods have also been proposed (*see* e.g., U.S. Patent No. 5,965,408 and the references above). In this approach, double stranded DNAs corresponding to one or more genes sharing regions of sequence similarity are combined and denatured, in the presence or absence of primers specific for the

gene. The single stranded polynucleotides are then annealed and incubated in the presence of a polymerase and a chain terminating reagent (e.g., ultraviolet, gamma or X-ray irradiation; ethidium bromide or other intercalators; DNA binding proteins, such as single strand binding proteins, transcription activating factors, or histones; polycyclic aromatic hydrocarbons; trivalent chromium or a trivalent chromium salt; or abbreviated polymerization mediated by rapid thermocycling; and the like), resulting in the production of partial duplex molecules. The partial duplex molecules, e.g., containing partially extended chains, are then denatured and reannealed in subsequent rounds of replication or partial replication resulting in polynucleotides which share varying degrees of sequence similarity and which are chimeric with respect to the starting population of DNA molecules. Optionally, the products or partial pools of the products can be amplified at one or more stages in the process. Polynucleotides produced by a chain termination method, such as described above are suitable substrates for DNA shuffling according to any of the described formats.

Diversity can be further increased by using non-homology based shuffling methods (which, as set forth in the above publications and applications can be homology or non-homology based, depending on the precise format). For example, incremental truncation for the creation of hybrid enzymes (ITCHY) described in Ostermeier et al. (1999) "A combinatorial approach to hybrid enzymes independent of DNA homology" Nature Biotech 17:1205, can be used to generate an initial a shuffled library which can optionally serve as a substrate for one or more rounds of in vitro or in vivo shuffling methods. See, also, Ostermeier et al. (1999) "Combinatorial Protein Engineering by Incremental Truncation," Proc. Natl. Acad. Sci. USA, 96: 3562-67; Ostermeier et al. (1999), "Incremental Truncation as a Strategy in the Engineering of Novel Biocatalysts," Biological and Medicinal Chemistry, 7: 2139-44.

Methods for generating multispecies expression libraries have been described (e.g., U.S. Patent Nos. 5,783,431; 5,824,485 and the references above) and their use to identify protein activities of interest has been proposed (U.S. Patent 5,958,672 and the references above). Multispecies expression libraries are, in general, libraries comprising cDNA or genomic sequences from a plurality of species or strains, operably linked to appropriate regulatory sequences, in an expression cassette. The cDNA and/or genomic sequences are optionally randomly concatenated to further enhance diversity. The vector can be a shuttle vector suitable for transformation and expression in more than one species

of host organism, e.g., bacterial species, eukaryotic cells. In some cases, the library is biased by preselecting sequences which encode a protein of interest, or which hybridize to a nucleic acid of interest. Any such libraries can be provided as substrates for any of the methods herein described.

In some applications, it is desirable to preselect or prescreen libraries (e.g., an amplified library, a genomic library, a cDNA library, a normalized library, etc.) or other substrate nucleic acids prior to shuffling, or to otherwise bias the substrates towards nucleic acids that encode functional products (shuffling procedures can also, independently have these effects). For example, in the case of antibody engineering, it is possible to bias the shuffling process toward antibodies with functional antigen binding sites by taking advantage of in vivo recombination events prior to DNA shuffling by any described method. For example, recombined CDRs derived from B cell cDNA libraries can be amplified and assembled into framework regions (e.g., Jirholt et al. (1998) "Exploiting sequence space: shuffling in vivo formed complementarity determining regions into a master framework" Gene 215: 471) prior to DNA shuffling according to any of the methods described herein.

Libraries can be biased towards nucleic acids which encode proteins with desirable enzyme activities. For example, after identifying a clone from a library which exhibits a specified activity, the clone can be mutagenized using any known method for introducing DNA alterations, including, but not restricted to, DNA shuffling. A library comprising the mutagenized homologues is then screened for a desired activity, which can be the same as or different from the initially specified activity. An example of such a procedure is proposed in U.S. Patent No. 5,939,250. Desired activities can be identified by any method known in the art. For example, WO 99/10539 proposes that gene libraries can be screened by combining extracts from the gene library with components obtained from metabolically rich cells and identifying combinations which exhibit the desired activity. It has also been proposed (e.g., WO 98/58085) that clones with desired activities can be identified by inserting bioactive substrates into samples of the library, and detecting bioactive fluorescence corresponding to the product of a desired activity using a fluorescent analyzer, e.g., a flow cytometry device, a CCD, a fluorometer, or a spectrophotometer.

Libraries can also be biased towards nucleic acids which have specified characteristics, e.g., hybridization to a selected nucleic acid probe. For example, application WO 99/10539 proposes that polynucleotides encoding a desired activity (e.g., an enzymatic

activity, for example: a lipase, an esterase, a protease, a glycosidase, a glycosyl transferase, a phosphatase, a kinase, an oxygenase, a peroxidase, a hydrolase, a hydratase, a nitrilase, a transaminase, an amidase or an acylase) can be identified from among genomic DNA sequences in the following manner. Single stranded DNA molecules from a population of genomic DNA are hybridized to a ligand-conjugated probe. The genomic DNA can be derived from either a cultivated or uncultivated microorganism, or from an environmental sample. Alternatively, the genomic DNA can be derived from a multicellular organism, or a tissue derived therefrom.

Second strand synthesis can be conducted directly from the hybridization probe used in the capture, with or without prior release from the capture medium or by a wide variety of other strategies known in the art. Alternatively, the isolated single-stranded genomic DNA population can be fragmented without further cloning and used directly in a shuffling format that employs a single-stranded template. Such single-stranded template shuffling formats are described, for example, in WO 98/27230, "Methods and Compositions for Polypeptide Engineering" by Patten et al.; USSN 60/186,482 filed March 2, 2000, "Single-Stranded Nucleic Acid Template-Mediated Recombination and Nucleic Acid Fragment Isolation" by Affholter; WO 00/00632, "Methods for Generating Highly Diverse Libraries" by Wagner et al.; and WO 00/09679, "Methods for Obtaining in Vitro Recombined Polynucleotide Sequence Banks and Resulting Sequences." In one such method the fragment population derived the genomic library(ies) is annealed with partial, or, often approximately full length ssDNA or RNA corresponding to the opposite strand. Assembly of complex chimeric genes from this population is the mediated by nuclease-base removal of non-hybridizing fragment ends, polymerization to fill gaps between such fragments and subsequent single stranded ligation. The parental strand can be removed by digestion (if RNA or uracil-containing), magnetic separation under denaturing conditions (if labeled in a manner conducive to such separation) and other available separation/purification methods. Alternatively, the parental strand is optionally co-purified with the chimeric strands and removed during subsequent screening and processing steps.

In one approach, single-stranded molecules are converted to double-stranded DNA (dsDNA) and the dsDNA molecules are bound to a solid support by ligand-mediated binding. After separation of unbound DNA, the selected DNA molecules are released from the support and introduced into a suitable host cell to generate a library enriched sequences

which hybridize to the probe. A library produced in this manner provides a desirable substrate for further shuffling using any of the shuffling reactions described herein.

It will further be appreciated that any of the above described techniques suitable for enriching a library prior to shuffling can be used to screen the products generated by the methods of DNA shuffling.

The shuffling of a single gene and the shuffling of a family of genes provide two of the most powerful methods available for improving and "migrating" (gradually changing the type of reaction, substrate or activity of a selected enzyme) the functions of biocatalysts. When shuffling a family of genes, homologous sequences, *e.g.*, from different species or chromosomal positions, are recombined. In single gene shuffling, a single sequence is mutated or otherwise altered and then recombined. These formats share some common principles.

The breeding procedure starts with at least two substrates that generally show substantial sequence identity to each other (*i.e.*, at least about 30%, 50%, 70%, 80% or 90% sequence identity), but differ from each other at certain positions. The difference can be any type of mutation, for example, substitutions, insertions and deletions. Often, different segments differ from each other in about 5-20 positions. For recombination to generate increased diversity relative to the starting materials, the starting materials must differ from each other in at least two nucleotide positions. That is, if there are only two substrates, there should be at least two divergent positions. If there are three substrates, for example, one substrate can differ from the second at a single position, and the second can differ from the third at a different single position. The starting DNA segments can be natural variants of each other, for example, allelic or species variants. The segments can also be from nonallelic genes showing some degree of structural and usually functional relatedness (*e.g.*, different genes within a superfamily, such as the arene dioxygenase super family). The starting DNA segments can also be induced variants of each other. For example, one DNA segment can be produced by error-prone PCR replication of the other, or by substitution of a mutagenic cassette. Induced mutants can also be prepared by propagating one (or both) of the segments in a mutagenic strain. In these situations, strictly speaking, the second DNA segment is not a single segment but a large family of related segments. The different segments forming the starting materials are often the same length or substantially the same length. However, this need not be the case; for example; one segment can be a subsequence

of another. The segments can be present as part of larger molecules, such as vectors, or can be in isolated form.

The starting DNA segments are recombined by any of the sequence recombination formats provided herein to generate a diverse library of recombinant DNA segments. Such a library can vary widely in size from having fewer than 10 to more than 10^5, 10^9, 10^{12} or more members. In some embodiments, the starting segments and the recombinant libraries generated will include full-length coding sequences and any essential regulatory sequences, such as a promoter and polyadenylation sequence, required for expression. In other embodiments, the recombinant DNA segments in the library can be inserted into a common vector providing sequences necessary for expression before performing screening/selection.

1. Use of Restriction Enzyme Sites to Recombine Mutations

In some situations it is advantageous to use restriction enzyme sites in nucleic acids to direct the recombination of mutations in a nucleic acid sequence of interest. These techniques are particularly preferred in the evolution of fragments that cannot readily be shuffled by existing methods due to the presence of repeated DNA or other problematic primary sequence motifs. These situations also include recombination formats in which it is preferred to retain certain sequences unmutated. The use of restriction enzyme sites is also preferred for shuffling large fragments (typically greater than 10 kb), such as gene clusters that cannot be readily shuffled and "PCR-amplified" because of their size. Although fragments up to 50 kb have been reported to be amplified by PCR (Barnes, *Proc. Natl. Acad. Sci. U.S.A.* **91**:2216-2220 (1994)), it can be problematic for fragments over 10 kb, and thus alternative methods for shuffling in the range of 10 - 50 kb and beyond are preferred. Preferably, the restriction endonucleases used are of the Class II type (Sambrook, Ausubel and Berger, *supra*) and of these, preferably those which generate nonpalindromic sticky end overhangs such as Alwn I, Sfi I or BstX1. These enzymes generate nonpalindromic ends that allow for efficient ordered reassembly with DNA ligase. Typically, restriction enzyme (or endonuclease) sites are identified by conventional restriction enzyme mapping techniques (Sambrook, Ausubel, and Berger, *supra.*), by analysis of sequence information for that gene, or by introduction of desired restriction sites into a nucleic acid sequence by synthesis (*i.e.* by incorporation of silent mutations).

The DNA substrate molecules to be digested can either be from *in vivo* replicated DNA, such as a plasmid preparation, or from PCR amplified nucleic acid

fragments harboring the restriction enzyme recognition sites of interest, preferably near the ends of the fragment. Typically, at least two variants of a gene of interest, each having one or more mutations, are digested with at least one restriction enzyme determined to cut within the nucleic acid sequence of interest. The restriction fragments are then joined with DNA ligase to generate full length genes having shuffled regions. The number of regions shuffled will depend on the number of cuts within the nucleic acid sequence of interest. The shuffled molecules can be introduced into cells as described above and screened or selected for a desired property as described herein. Nucleic acids can then be isolated from pools (libraries), or clones having desired properties and subjected to the same procedure until a desired degree of improvement is obtained.

In some embodiments, at least one DNA substrate molecule or fragment thereof is isolated and subjected to mutagenesis. In some embodiments, the pool or library of religated restriction fragments are subjected to mutagenesis before the digestion-ligation process is repeated. "Mutagenesis" as used herein includes such techniques known in the art as PCR mutagenesis, oligonucleotide-directed mutagenesis, site-directed mutagenesis, *etc.*, and recursive sequence recombination by any of the techniques described herein.

2. *Reassembly PCR*

A further technique for recombining mutations in a nucleic acid sequence utilizes "reassembly PCR." This method can be used to assemble multiple segments that have been separately evolved into a full length nucleic acid template such as a gene. This technique is performed when a pool of advantageous mutants is known from previous work or has been identified by screening mutants that may have been created by any mutagenesis technique known in the art, such as PCR mutagenesis, cassette mutagenesis, doped oligo mutagenesis, chemical mutagenesis, or propagation of the DNA template *in vivo* in mutator strains. Boundaries defining segments of a nucleic acid sequence of interest preferably lie in intergenic regions, introns, or areas of a gene not likely to have mutations of interest. Preferably, oligonucleotide primers (oligos) are synthesized for PCR amplification of segments of the nucleic acid sequence of interest, such that the sequences of the oligonucleotides overlap the junctions of two segments. The overlap region is typically about 10 to 100 nucleotides in length. Each of the segments is amplified with a set of such primers. The PCR products are then "reassembled" according to assembly protocols such as those discussed herein to assemble randomly fragmented genes. In brief, in an assembly protocol the PCR products are first purified away from the primers, by, for example, gel

electrophoresis or size exclusion chromatography. Purified products are mixed together and subjected to about 1-10 cycles of denaturing, reannealing, and extension in the presence of polymerase and deoxynucleoside triphosphates (dNTP's) and appropriate buffer salts in the absence of additional primers ("self-priming"). Subsequent PCR with primers flanking the gene are used to amplify the yield of the fully reassembled and shuffled genes.

In some embodiments, the resulting reassembled genes are subjected to mutagenesis before the process is repeated.

In a further embodiment, the PCR primers for amplification of segments of the nucleic acid sequence of interest are used to introduce variation into the gene of interest as follows. Mutations at sites of interest in a nucleic acid sequence are identified by screening or selection, by sequencing homologues of the nucleic acid sequence, and so on. Oligonucleotide PCR primers are then synthesized which encode wild type or mutant information at sites of interest. These primers are then used in PCR mutagenesis to generate libraries of full length genes encoding permutations of wild type and mutant information at the designated positions. This technique is typically advantageous in cases where the screening or selection process is expensive, cumbersome, or impractical relative to the cost of sequencing the genes of mutants of interest and synthesizing mutagenic oligonucleotides.

3. Site Directed Mutagenesis (SDM) with Oligonucleotides Encoding Homologue Mutations Followed by Shuffling

In some embodiments of the invention, sequence information from one or more substrate sequences is added to a given "parental" sequence of interest, with subsequent recombination between rounds of screening or selection. Typically, this is done with site-directed mutagenesis performed by techniques well known in the art (*e.g.*, Berger, Ausubel and Sambrook, *supra.*) with one substrate as template and oligonucleotides encoding single or multiple mutations from other substrate sequences, *e.g.* homologous genes. After screening or selection for an improved phenotype of interest, the selected recombinant(s) can be further evolved using RSR techniques described herein. After screening or selection, site-directed mutagenesis can be done again with another collection of oligonucleotides encoding homologue mutations, and the above process repeated until the desired properties are obtained.

When the difference between two homologues is one or more single point mutations in a codon, degenerate oligonucleotides can be used that encode the sequences in

both homologues. One oligonucleotide can include many such degenerate codons and still allow one to exhaustively search all permutations over that block of sequence.

When the homologue sequence space is very large, it can be advantageous to restrict the search to certain variants. Thus, for example, computer modeling tools (Lathrop *et al., J. Mol. Biol.* **255**:641-665 (1996)) can be used to model each homologue mutation onto the target protein and discard any mutations that are predicted to grossly disrupt structure and function.

4. In vitro Nucleic Acid Shuffling Formats

In one embodiment for shuffling DNA sequences *in vitro*, the initial substrates for recombination are a pool of related sequences, *e.g.*, different variant forms, as homologs from different individuals, strains, or species of an organism, or related sequences from the same organism, as allelic variations. The sequences can be DNA or RNA and can be of various lengths depending on the size of the gene or DNA fragment to be recombined or reassembled. Preferably the sequences are from 50 base pairs (bp) to 50 kilobases (kb).

The pool of related substrates are converted into overlapping fragments, *e.g.*, from about 5 bp to 5 kb or more. Often, for example, the size of the fragments is from about 10 bp to 1000 bp, and sometimes the size of the DNA fragments is from about 100 bp to 500 bp. The conversion can be effected by a number of different methods, such as DNase I or RNase digestion, random shearing or partial restriction enzyme digestion. For discussions of protocols for the isolation, manipulation, enzymatic digestion, and the like of nucleic acids, see, for example, Sambrook *et al.* and Ausubel, both *supra*. The concentration of nucleic acid fragments of a particular length and sequence is often less than 0.1 % or 1% by weight of the total nucleic acid. The number of different specific nucleic acid fragments in the mixture is usually at least about 100, 500 or 1000.

The mixed population of nucleic acid fragments are converted to at least partially single-stranded form using a variety of techniques, including, for example, heating, chemical denaturation, use of DNA binding proteins, and the like. Conversion can be effected by heating to about 80 °C to 100 °C, more preferably from 90 °C to 96 °C, to form single-stranded nucleic acid fragments and then reannealing. Conversion can also be effected by treatment with single-stranded DNA binding protein (see Wold, *Annu. Rev. Biochem.* **66**:61-92 (1997)) or *rec*A protein (*see, e.g.*, Kiianitsa, *Proc. Natl. Acad. Sci. USA* **94**:7837-7840 (1997)). Single-stranded nucleic acid fragments having regions of sequence

identity with other single-stranded nucleic acid fragments can then be reannealed by cooling to 20 °C to 75 °C, and preferably from 40 °C to 65 °C. Renaturation can be accelerated by the addition of polyethylene glycol (PEG), other volume-excluding reagents or salt. The salt concentration is preferably from 0 mM to 200 mM, more preferably the salt concentration is from 10 mM to 100 mM. The salt may be KCl or NaCl. The concentration of PEG is preferably from 0% to 20%, more preferably from 5% to 10%. The fragments that reanneal can be from different substrates. The annealed nucleic acid fragments are incubated in the presence of a nucleic acid polymerase, such as Taq or Klenow, and dNTP's (*i.e.* dATP, dCTP, dGTP and dTTP). If regions of sequence identity are large, Taq polymerase can be used with an annealing temperature of between 45-65 °C. If the areas of identity are small, Klenow polymerase can be used with an annealing temperature of between 20-30 °C. The polymerase can be added to the random nucleic acid fragments prior to annealing, simultaneously with annealing or after annealing.

The process of denaturation, renaturation and incubation in the presence of polymerase of overlapping fragments to generate a collection of polynucleotides containing different permutations of fragments is sometimes referred to as shuffling of the nucleic acid *in vitro*. This cycle is repeated for a desired number of times. Preferably the cycle is repeated from 2 to 100 times, more preferably the sequence is repeated from 10 to 40 times. The resulting nucleic acids are a family of double-stranded polynucleotides of from about 50 bp to about 100 kb, preferably from 500 bp to 50 kb. The population represents variants of the starting substrates showing substantial sequence identity thereto but also diverging at several positions. The population has many more members than the starting substrates. The population of fragments resulting from shuffling is used to transform host cells, optionally after cloning into a vector.

In one embodiment utilizing *in vitro* shuffling, subsequences of recombination substrates can be generated by amplifying the full-length sequences under conditions which produce a substantial fraction, typically at least 20 percent or more, of incompletely extended amplification products. Another embodiment uses random primers to prime the entire template DNA to generate less than full length amplification products. The amplification products, including the incompletely extended amplification products are denatured and subjected to at least one additional cycle of reannealing and amplification. This variation, in which at least one cycle of reannealing and amplification provides a

substantial fraction of incompletely extended products, is termed "stuttering." In the subsequent amplification round, the partially extended (less than full length) products reanneal to and prime extension on different sequence-related template species. In another embodiment, the conversion of substrates to fragments can be effected by partial PCR amplification of substrates.

In another embodiment, a mixture of fragments is spiked with one or more oligonucleotides. The oligonucleotides can be designed to include precharacterized mutations of a wildtype sequence, or sites of natural variations between individuals or species. The oligonucleotides also include sufficient sequence or structural homology flanking such mutations or variations to allow annealing with the wildtype fragments. Annealing temperatures can be adjusted depending on the length of homology.

In a further embodiment, recombination occurs in at least one cycle by template switching, such as when a DNA fragment derived from one template primes on the homologous position of a related but different template. Template switching can be induced by addition of recA (*see*, Kiianitsa (1997) *supra*), rad51 (*see*, Namsaraev, *Mol. Cell. Biol.* **17**:5359-5368 (1997)), rad55 (*see*, Clever, *EMBO J.* **16**:2535-2544 (1997)), rad57 (*see*, Sung, *Genes Dev.* **11**:1111-1121 (1997)) or other polymerases (*e.g.*, viral polymerases, reverse transcriptase) to the amplification mixture. Template switching can also be increased by increasing the DNA template concentration.

Another embodiment utilizes at least one cycle of amplification, which can be conducted using a collection of overlapping single-stranded DNA fragments of related sequence, and different lengths. Fragments can be prepared using a single stranded DNA phage, such as M13 (*see*, Wang, *Biochemistry* **36**:9486-9492 (1997)). Each fragment can hybridize to and prime polynucleotide chain extension of a second fragment from the collection, thus forming sequence-recombined polynucleotides. In a further variation, ssDNA fragments of variable length can be generated from a single primer by Pfu, Taq, Vent, Deep Vent, UlTma DNA polymerase or other DNA polymerases on a first DNA template (*see*, Cline, *Nucleic Acids Res.* **24**:3546-3551 (1996)). The single stranded DNA fragments are used as primers for a second, Kunkel-type template, consisting of a uracil-containing circular ssDNA. This results in multiple substitutions of the first template into the second. *See*, Levichkin, *Mol. Biology* **29**:572-577 (1995); Jung, *Gene* **121**:17-24 (1992).

In some embodiments of the invention, shuffled nucleic acids obtained by use of the recursive recombination methods of the invention, are put into a cell and/or

organism for screening. Shuffled dioxygenase genes can be introduced into, for example, bacterial cells (including cyanobacteria), yeast cells, fungal cells, vertebrate cells, invertebrate cells or plant cells for initial screening. Bacterial species, such as *E. coli, Pseudomonas sp, Bacillus. subtilis, Burkholderia cepacia, Alcaligenes, Acinetobacter, Rhodococcus Arthrobacter, Sphingomonas* are preferred examples of suitable bacterial cells into which one can insert and express shuffled dioxygenase genes which provide for convenient shuttling to other cell types (a variety of vectors for shuttling material between these bacterial cells and eukaryotic cells are available; *see,* Sambrook, Ausubel and Berger, *all supra*). The shuffled genes can be introduced into bacterial, fungal or yeast cells either by integration into the chromosomal DNA or as plasmids.

Although bacterial and yeast systems are most preferred in the present invention, in one embodiment, shuffled genes can also be introduced into plant cells for production purposes (it will be appreciated that transgenic plants are, increasingly, an important source of industrial enzymes). Thus, a transgene of interest can be modified using the recursive sequence recombination methods of the invention *in vitro* and reinserted into the cell for *in vivo/in situ* selection for the new or improved dioxygenase property, in bacteria, eukaryotic cells, or whole eukaryotic organisms.

5. In vivo Nucleic Acid Shuffling Formats

In some embodiments of the invention, DNA substrate molecules are introduced into cells, wherein the cellular machinery directs their recombination. For example, a library of mutants is constructed and screened or selected for mutants with improved phenotypes by any of the techniques described herein. The DNA substrate molecules encoding the best candidates are recovered by any of the techniques described herein, then fragmented and used to transfect a plant host and screened or selected for improved function. If further improvement is desired, the DNA substrate molecules are recovered from the host cell, such as by PCR, and the process is repeated until a desired level of improvement is obtained. In some embodiments, the fragments are denatured and reannealed prior to transfection, coated with recombination stimulating proteins such as recA, or co-transfected with a selectable marker such as Neo^R to allow the positive selection for cells receiving recombined versions of the gene of interest. Methods for *in vivo* shuffling are described in, for example, PCT application WO 98/13487 and WO 97/20078.

The efficiency of *in vivo* shuffling can be enhanced by increasing the copy number of a gene of interest in the host cells. For example, the majority of bacterial cells in

stationary phase cultures grown in rich media contain two, four or eight genomes. In minimal medium the cells contain one or two genomes. The number of genomes per bacterial cell thus depends on the growth rate of the cell as it enters stationary phase. This is because rapidly growing cells contain multiple replication forks, resulting in several genomes in the cells after termination. The number of genomes is strain dependent, although all strains tested have more than one chromosome in stationary phase. The number of genomes in stationary phase cells decreases with time. This appears to be due to fragmentation and degradation of entire chromosomes, similar to apoptosis in mammalian cells. This fragmentation of genomes in cells containing multiple genome copies results in massive recombination and mutagenesis. The presence of multiple genome copies in such cells results in a higher frequency of homologous recombination in these cells, both between copies of a gene in different genomes within the cell, and between a genome within the cell and a transfected fragment. The increased frequency of recombination allows one to evolve a gene more quickly to acquire optimized characteristics.

In nature, the existence of multiple genomic copies in a cell type would usually not be advantageous due to the greater nutritional requirements needed to maintain this copy number. However, artificial conditions can be devised to select for high copy number. Modified cells having recombinant genomes are grown in rich media (in which conditions, multicopy number should not be a disadvantage) and exposed to a mutagen, such as ultraviolet or gamma irradiation or a chemical mutagen, *e.g.*, mitomycin, nitrous acid, photoactivated psoralens, alone or in combination, which induces DNA breaks amenable to repair by recombination. These conditions select for cells having multicopy number due to the greater efficiency with which mutations can be excised. Modified cells surviving exposure to mutagen are enriched for cells with multiple genome copies. If desired, selected cells can be individually analyzed for genome copy number (*e.g.*, by quantitative hybridization with appropriate controls). For example, individual cells can be sorted using a cell sorter for those cells containing more DNA, *e.g.*, using DNA specific fluorescent compounds or sorting for increased size using light dispersion. Some or all of the collection of cells surviving selection are tested for the presence of a gene that is optimized for the desired property.

In one embodiment, phage libraries are made and recombined in mutator strains such as cells with mutant or impaired gene products of mutS, mutT, mutH, mutL, ovrD, dcm, vsr, umuC, umuD, sbcB, recJ, *etc.* The impairment is achieved by genetic

mutation, allelic replacement, selective inhibition by an added reagent such as a small compound or an expressed antisense RNA, or other techniques. High multiplicity of infection (MOI) libraries are used to infect the cells to increase recombination frequency.

Additional strategies for making phage libraries and or for recombining DNA from donor and recipient cells are set forth in U.S. Pat. No. 5,521,077. Additional recombination strategies for recombining plasmids in yeast are set forth in WO 97 07205.

6. Whole Genome Shuffling

In one embodiment, the selection methods herein are utilized in a "whole genome shuffling" format. An extensive guide to the many forms of whole genome shuffling is found in the pioneering application to the inventors and their co-workers entitled "Evolution of Whole Cells and Organisms by Recursive Sequence Recombination," Attorney Docket No. 018097-020720US filed July 15, 1998 by del Cardayre *et al.* (USSN 09/161,188).

In brief, whole genome shuffling makes no presuppositions at all regarding what nucleic acids may confer a desired property. Instead, entire genomes (*e.g.*, from a genomic library, or isolated from an organism) are shuffled in cells and selection protocols applied to the cells.

The fermentation of microorganisms for the production of natural products is the oldest and most sophisticated application of biocatalysis.

The methods herein allow dioxygenase biocatalysts to be improved at a faster pace than conventional methods. Whole genome shuffling can at least double the rate of strain improvement for microorganisms used in fermentation as compared to traditional methods. This provides for a relative decrease in the cost of fermentation processes. New products can enter the market sooner, producers can increase profits as well as market share, and consumers gain access to more products of higher quality and at lower prices. Further, increased efficiency of production processes translates to less waste production and more frugal use of resources. Whole genome shuffling provides a means of accumulating multiple useful mutation per cycle and thus eliminate the inherent limitation of current strain improvement programs (SIPs).

Nucleic acid shuffling provides recursive mutagenesis, recombination, and selection of DNA sequences. A key difference between nucleic acid shuffling-mediated recombination and natural sexual recombination is that nucleic acid shuffling effects both the pairwise (two parents) and the poolwise (multiple parents) recombination of parent

molecules. Natural recombination is more conservative and is limited to pairwise recombination. In nature, pairwise recombination provides stability within a population by preventing large leaps in sequences or genomic structure that can result from poolwise recombination. However, for the purposes of directed evolution, poolwise recombination is appealing since the beneficial mutations of multiple parents can be combined during a single cross to produce a superior offspring. Poolwise recombination is analogous to the crossbreeding of inbred strains in classic strain improvement, except that the crosses occur between many strains at once. In essence, poolwise recombination is a sequence of events that effects the recombination of a population of nucleic acid sequences that results in the generation of new nucleic acids that contains genetic information from more than two of the original nucleic acids.

There are a few general methods for effecting efficient recombination in prokaryotes. Bacteria have no known sexual cycle *per se*, but there are natural mechanisms by which the genomes of these organisms undergo recombination. These mechanisms include natural competence, phage-mediated transduction, and cell-cell conjugation. Bacteria that are naturally competent are capable of efficiently taking up naked DNA from the environment. If homologous, this DNA undergoes recombination with the genome of the cell, resulting in genetic exchange. *Bacillus subtilis*, the primary production organism of the enzyme industry, is known for the efficiency with which it carries out this process.

In generalized transduction, a bacteriophage mediates genetic exchange. A transducing phage will often package headfulls of the host genome. These phage can infect a new host and deliver a fragment of the former host genome which is frequently integrated via homologous recombination. Cells can also transfer DNA between themselves by conjugation. Cells containing the appropriate mating factors transfer episomes as well as entire chromosomes to an appropriate acceptor cell where it can recombine with the acceptor genome. Conjugation resembles sexual recombination for microbes and can be intraspecific, interspecific, and intergeneric. For example, an efficient means of transforming *Streptomyces* sp., a genera responsible for producing many commercial antibiotics, is by the conjugal transfer of plasmids from *Echerichia coli*.

For many industrial microorganisms, knowledge of competence, transducing phage, or fertility factors is lacking. Protoplast fusion has been developed as a versatile and general alternative to these natural methods of recombination. Protoplasts are prepared by removing the cell wall by treating cells with lytic enzymes in the presence of osmotic

stabilizers. In the presence of a fusogenic agent, such as polyethylene glycol (PEG), protoplasts are induced to fuse and form transient hybrids or "fusants." During this hybrid state, genetic recombination occurs at high frequency allowing the genomes to reassort. The final step is the successful segregation and regeneration of viable cells from the fused protoplasts. Protoplast fusion can be intraspecific, interspecific, and intergeneric and has been applied to both prokaryotes and eukaryotes. In addition, it is possible to fuse more than two cells, thus providing a mechanism for effecting poolwise recombination. While no fertility factors, transducing phages or competency development is needed for protoplast fusion, a method for the formation, fusing, and regeneration of protoplasts is typically optimized for each organism.

Modifications can be made to the method and materials as hereinbefore described without departing from the spirit or scope of the invention as claimed, and the invention can be put to a number of different uses, including:

The use of an integrated system to test dioxygenase in shuffled DNAs, including in an iterative process.

7. Shuffling families of arene dioxygenases

For identifying homologous genes used to shuffle a family of genes, representative alignments of monooxygenase enzymes can be found in the literature or generated from sequences retrieved from GeneBank or an associated public database

(a). Target sequences for nucleic acid shuffling of ADOs and variations

The following discussion focuses on ADO for clarity of illustration. Those of skill in the art will recognize that this discussion is illustrative of DOs in general, and is not limited to ADOs.

Because all known ADOs are multicomponent enzymes having from 2 to 4 functionally different subunits, any of the sequence components comprising a particular ADO or a family of homologous ADOs can be shuffled. For the purposes of evolving ADOs with changes in specificity, regioselectivity and mode of action (*e.g.* dioxygenase or monooxygenase activity) for oxidizing a particular compound of interest, it is often preferred to shuffle nucleic acids encoding the iron-sulfur protein of terminal oxygenase (where terminal oxygenase is made of two polypeptides, the large subunit is the preferred target sequence for shuffling).

For the purpose of improving overall turnover rates and efficiency of electron transfer between all components of dioxygenases, shuffling of the polynucleotide

sequences encoding other functional polypeptides (reductase, ferredoxin, small terminal oxidase component) is a preferred embodiment. These sequences can be shuffled individually, in sub-sets or as a part of a gene cluster which encodes all of the ADO polypeptides. This allows for changes in both the coding polynucleotide sequences, and also for generating combinations of functional chimeric ADOs with various functions assembled from two or more parental sequences (family gene cluster shuffling).

Many natural dioxygenases are in effect fusion proteins where the functions of reductase and ferredoxin are combined in one polypeptide. Moreover, the terminal oxygenase component is presented by one polypeptide instead of two. For one skilled in the art, it is apparent that polynucleotides encoding natural and artificial fusions of ADO components can be also used for shuffling. For example, in a recursive shuffling process, in principle, all desired characteristics of an ADO can be altered and improved in the desired directions regardless of which particular polynucleotide subsequence is selected from a complete ADO coding sequence.

To illustrate shuffling a family of genes to improve arene dioxygenase enzymes, one or more of the more than 50 members of this superfamily is selected, aligned with similar homologous sequences, shuffled against these homologous sequences and screened.

The screening is done most easily in a bacterial system. DNA from clones with improved activity can be shuffled together in subsequent rounds of shuffling and screened for further improvement.

8. Codon Modification Shuffling

Procedures for codon modification shuffling are described in detail in SHUFFLING OF CODON ALTERED GENES, Phillip A. Patten and Willem P.C. Stemmer, filed September 29, 1998, USSN 60/102362 and in SHUFFLING OF CODON ALTERED GENES, Phillip A. Patten and Willem P.C. Stemmer, filed January 29, 1999, USSN 60/117729. In brief, by synthesizing nucleic acids in which the codons encoding polypeptides are altered, it is possible to access a completely different mutational cloud upon subsequent mutation of the nucleic acid. This increases the sequence diversity of the starting nucleic acids for shuffling protocols, which alters the rate and results of forced evolution procedures. Codon modification procedures can be used to modify any nucleic acid described herein, *e.g.*, prior to performing nucleic acid shuffling, or codon modification

approaches can be used in conjunction with oligonucleotide shuffling procedures as described *supra*.

In these methods, a first nucleic acid sequence encoding a first polypeptide sequence is selected. A plurality of codon altered nucleic acid sequences, each of which encode the first polypeptide, or a modified or related polypeptide, is then selected (*e.g.*, a library of codon altered nucleic acids can be selected in a biological assay which recognizes library components or activities), and the plurality of codon-altered nucleic acid sequences is recombined to produce a target codon altered nucleic acid encoding a second protein. The target codon altered nucleic acid is then screened for a detectable functional or structural property, optionally including comparison to the properties of the first polypeptide and/or related polypeptides. The goal of such screening is to identify a polypeptide that has a structural or functional property equivalent or superior to the first polypeptide or related polypeptide. A nucleic acid encoding such a polypeptide can be used in essentially any procedure desired, including introducing the target codon altered nucleic acid into a cell, vector, virus, attenuated virus (*e.g.*, as a component of a vaccine or immunogenic composition), transgenic organism, or the like.

9. Oligonucleotide and in silico shuffling formats

In addition to the formats for shuffling noted above, at least two additional related formats are useful in the practice of the present invention. The first, referred to as "in silico" shuffling utilizes computer algorithms to perform "virtual" shuffling using genetic operators in a computer. As applied to the present invention, gene sequence strings are recombined in a computer system and desirable products are made, *e.g.*, by reassembly PCR of synthetic oligonucleotides. In silico shuffling is described in detail in Selifonov and Stemmer in "METHODS FOR MAKING CHARACTER STRINGS, POLYNUCLEOTIDES & POLYPEPTIDES HAVING DESIRED CHARACTERISTICS" filed February 5, 1999, USSN 60/118854. In brief, genetic operators (algorithms which represent given genetic events such as point mutations, recombination of two strands of homologous nucleic acids, *etc.*) are used to model recombinational or mutational events which can occur in one or more nucleic acid, *e.g.*, by aligning nucleic acid sequence strings (using standard alignment software, or by manual inspection and alignment) and predicting recombinational outcomes. The predicted recombinational outcomes are used to produce corresponding molecules, *e.g.*, by oligonucleotide synthesis and reassembly PCR.

The second useful format is referred to as "oligonucleotide mediated shuffling" in which oligonucleotides corresponding to a family of related homologous nucleic acids (*e.g.*, as applied to the present invention, interspecific or allelic variants of a dioxygenase nucleic acid) which are recombined to produce selectable nucleic acids. This format is described in detail in Crameri *et al.* "OLIGONUCLEOTIDE MEDIATED NUCLEIC ACID RECOMBINATION" filed February 5, 1999, USSN 60/118,813 and Crameri et al. "OLIGONUCLEOTIDE MEDIATED NUCLEIC ACID RECOMBINATION" filed June 24, 1999, USSN 60/141,049. The technique can be used to recombine homologous or even non-homologous nucleic acid sequences.

One advantage of the oligonucleotide-mediated recombination is the ability to recombine homologous nucleic acids with low sequence similarity, or even non-homologous nucleic acids. In these low-homology oligonucleotide shuffling methods, one or more set of fragmented nucleic acids are recombined, *e.g.*, with a with a set of crossover family diversity oligonucleotides. Each of these crossover oligonucleotides have a plurality of sequence diversity domains corresponding to a plurality of sequence diversity domains from homologous or non-homologous nucleic acids with low sequence similarity. The fragmented oligonucleotides, which are derived by comparison to one or more homologous or non-homologous nucleic acids, can hybridize to one or more region of the crossover oligos, facilitating recombination.

When recombining homologous nucleic acids, sets of overlapping family gene oligonucleotides (which are derived by comparison of homologous nucleic acids and synthesis of oligonucleotide fragments) are hybridized and elongated (*e.g.*, by reassembly PCR), providing a population of recombined nucleic acids, which can be selected for a desired trait or property. Typically, the set of overlapping family genes include a plurality of oligonucleotide member types which have consensus region subsequences derived from a plurality of homologous target nucleic acids.

Typically, family gene shuffling oligonucleotides are provided by aligning homologous nucleic acid sequences to select conserved regions of sequence identity and regions of sequence diversity. A plurality of family gene shuffling oligonucleotides are synthesized (serially or in parallel) which correspond to at least one region of sequence diversity.

Sets of fragments, or subsets of fragments, used in oligonucleotide shuffling approaches can be provided by cleaving one or more homologous nucleic acids (*e.g.*, with a

DNase), or, more commonly, by synthesizing a set of oligonucleotides corresponding to a plurality of regions of at least one nucleic acid (typically oligonucleotides corresponding to a full-length nucleic acid are provided as members of a set of nucleic acid fragments). In the shuffling procedures herein, these cleavage fragments (*e.g.*, fragments of dioxygenases) can be used in conjunction with family gene shuffling oligonucleotides, *e.g.*, in one or more recombination reaction to produce recombinant dioxygenase nucleic acids.

10. Chimeric shuffling templates

In addition to the naturally occurring, mutated and synthetic oligonucleotides discussed above, polynucleotides encoding chimeric polypeptides can be used as substrates for shuffling in any of the above-described shuffling formats. Preferred chimeras have a shuffled active site or a shuffled active site region. Art-recognized methods for preparing chimeras are applicable to the methods described herein (*see*, for example, Shimoji *et al.*, *Biochemistry* **37**: 8848-8852 (1998)).

B. Reactions of Improved Dioxygenases

In another aspect, the invention provides a method for obtaining a polynucleotide encoding an improved dioxygenase polypeptide acting on an organic substrate. Presently preferred substrates include a target group selected from classes of substrates shown in Figures 14A-14R. The improved polypeptide exhibits one or more improved properties, compared to a naturally occurring polypeptide acting on the substrate(s). The method involves: (a) creating a library of recombinant polynucleotides encoding a dioxygenase polypeptide acting on the substrate; and (b) screening the library to identify a recombinant polynucleotide encoding an improved polypeptide that exhibits one or more improved properties compared to a naturally occurring dioxygenase polypeptide.

In a preferred embodiment, the library of recombinant polynucleotides is created by recombining at least a first form and a second form of a nucleic acid. At least one of these forms encodes the naturally occurring polypeptide or a fragment thereof. Preferably, the first form and the second form differ from each other in two or more nucleotides. In a further preferred embodiment, the first and second forms of the nucleic acid are homologous.

In addition to the methods described above for producing the encoding polynucleotides, the present invention also provides the polypeptides encoded by these

polynucleotides and methods of using these peptides for synthesizing valuable organic compounds. Some of these polypeptides and methods of using them are set forth below.

It is noted that the basic chemistry described below with reference to oxygenases in general and dioxygenases in particular is known. In addition to Ortiz de Montellano, *supra*, general guides to dioxygenase product chemistry include S. Brown and T. Hudlicky, in ORGANIC SYNTHESIS. THEORY AND APPLICATIONS (T.Hudlicky, Ed.,), volume 2, p. 113-176, JAI Press, Greenwich Connecticut (1993). Moreover, the various chemistries involved are found in Stryer BIOCHEMISTRY, third edition (or later editions) Freeman and Co. New York, NY (1988); Pine *et al.* ORGANIC CHEMISTRY FOURTH_EDITION (1980) McGraw-Hill, Inc. (USA) (or later editions); March, ADVANCED_ORGANIC CHEMISTRY REACTIONS, MECHANISMS and Structure 4th ed J. Wiley and Sons (New York, NY, 1992) (or later editions); Greene, *et al.*, PROTECTIVE GROUPS IN ORGANIC CHEMISTRY, 2nd Ed., John Wiley & Sons, New York, NY, 1991 (or later editions); Lide (ed) THE CRC HANDBOOK OF CHEMISTRY AND PHYSICS 75TH EDITION (1995) (or later editions); and in the references cited in the foregoing. Furthermore, an extensive guide to many chemical and industrial processes applicable to the present invention is found in the KIRK-OTHMER ENCYCLOPEDIA OF CHEMICAL TECHNOLOGY (third edition and fourth edition, through year 1998), Martin Grayson, Executive Editor, Wiley-Interscience, John Wiley and Sons, NY, and in the references cited therein ("Kirk-Othmer").

The following chemistries illustrate those generally accessible through the arene dioxygenase superfamily.

Shuffling approaches, such as shuffling a family of genes, apply to enhancing performance of dioxygenase polypeptides useful in each of the following classes of industrial chemical transformation. Other dioxygenase enzyme classes are also useful in practicing the present invention. Moreover, other polypeptides accessible through the present invention, and method of using these polypeptides will be apparent to those of skill in the art.

1. Oxidation of π-bonds to diols

In a preferred embodiment, the present invention provides improved polypeptides that can mediate the oxidation of π-bonds to vicinal diols. Also provided are host organisms expressing such improved polypeptides and methods of using these polypeptides and organisms in synthetic processes.

Among the enzymes known to oxidize π-bonds to the corresponding vicinal diols are the bacterial arene dioxygenases (ADOs). In the presence of oxygen, and of a reducing compound such as NAD(P)H, these enzymes catalyze the reductive dioxygenation of compounds as diverse as aromatic rings and non-aromatic multiple bonds.

Arene dioxygenases, such as toluene 2,3-dioxygenase, isopropylbenzene 2,3-dioxygenase, benzene-1,2-dioxygenase, biphenyl-2,3-dioxygenase, naphthalene-1,2-dioxygenase, and many homologous and/or functionally similar enzymes, constitute members of a class of enzymes useful in the manufacture of vicinal diols in a highly regioselective fashion. Moreover, the action of this class of enzymes on aromatic substrates does not involve formation of reactive arene epoxides or phenols. While potentially interesting from an academic standpoint, these enzymes have not been generally utilized due to several shortcomings. For example, these enzymes do not exhibit sufficient turnover numbers nor are they known to provide satisfactory regioselectivity for dihydroxylation of π-bonds in a substrates having more than one π-bond, or with more than one type of π-bond (*e.g.*, styrene).

Arene dioxygenases of various specificity, regioselectivity and enantiospecificity are capable of forming vicinal diols from a large array of substituted aromatic compounds and non-aromatic alkenes. For example, toluene dioxygenase has recently been implicated in dihydroxylation of several non-aromatic alkenes with concomitant formation of the glycol compounds of known and unknown stereochemistry. (Lange and Wackett, *J Bacteriol.*,**179**(12):3858-3865 (1997)). Similarly, purified naphthalene dioxygenase has been shown to catalyze dihydroxylation of styrene to (R)1-phenyl-1,2-ethanediol with enantiomeric excess of about 79% (Lee *et al.*, *Appl. Environ. Microbiol.*, **62**(9):3101-3106 (1996); Lee *et al.*, *J. Bacteriol.*, **178**(11):3353-3356 (1996)).

The non-phenolic nature of ring *cis*-dihydroxylation products arising from action of arene dioxygenases offers a significant advantage for manufacturing diols and other compounds by avoiding the accumulation toxic and reactive epoxide intermediates which may significantly impair the performance of the biocatalyst.

a. Oxidation of exocyclic and acyclic π-bonds

The present invention also provides improved polypeptides capable of oxidizing exocyclic and acyclic π-bonds for producing oxygen-containing species. For clarity of illustration, the following discussion focuses on the oxidation of olefins. This

focus is intended to be illustrative and not limiting of the scope of the invention. Many other appropriate substrates for oxidation using the methods of the invention will be apparent to those of skill in the art.

An exemplary oxidation of an olefin to the corresponding vicinal diol uses a dioxygenase to obtain the glycol directly from the olefin. This is best accomplished by recruiting a dioxygenase, such as an arene dioxygenase.

Arene dioxygenases are multi-component enzymes for which substrate specificity is primarily determined by the non-heme iron-sulfur cluster containing terminal oxidase protein(s), and, in the cases where terminal oxidase is comprised of two proteins, by the large subunit of the terminal oxidase. Other proteins such as ferredoxins and ferredoxin reductases provide transfer of electrons from reducing equivalents such as NAD(P)H to the terminal oxidase. Accordingly, where it is required to obtain a significant change in the substrate specificity, as compared to specificities of known arene dioxygenases existing in nature, nucleic acids that encode the terminal oxidase component(s) are the preferred substrate for the recombination and selection methods of the invention. Upon producing a library of recombinant polynucleotides as described herein, one can then select to identify those polynucleotides that encode an enzyme that has the desired change in substrate specificity. Such changes in the substrate specificity can include, for example, a gain of turnover of a novel substrate, a change in regioselectivity of oxidation, and a change in chirality of product formed. If the goal is primarily to obtain enzymes that have an increased catalytic turnover with an already acceptable substrate, shuffling of the nucleic acids encoding all of the components of arene dioxygenase is preferred.

Many arene dioxygenase genes that can be used as substrates for the recombination and selection methods of the invention are described in the art. Suitable arene dioxygenase-encoding polynucleotides can be obtained from many organisms using cloning methods known to one skilled in the art. The following list provides examples of polynucleotides that encode arene dioxygenases and are suitable for use in the methods of the invention. The loci are identified by GenBank ID and encode complete or partial protein components of the arene dioxygenases. Suitable loci include:

[PSETODC1C] toluene-1,2-dioxygenase;

[AF006691], [PJU53507], [PSECUMA], [REU24277] isopropylbenzene-2,3-[E04215], [PSEBDO] dioxygenase; benzene-1,2-dioxygenase; [U78099] tetrachlorobenzene dioxygenase; [AEBPHA1F], [CTU47637], [D78322], [D88020], [D88021], [PSEBPHA], [PSEBPHABC], [PSEBPHABCC], [PSU95054], [RERBPHA1], [RGBPHA], [RSU27591] biphenyl-2,3-dioxygenase; [PSU15298] chlorobenzene dioxygenase; [AB004059], [AF010471], [AF036940], [AF053735], [AF053736], pAF079317], [AF004283], [AF004284], [PSENAPDOXA], [PSENAPDOXB], [PSENDOABC], [PSEORF1], [PSU49496] naphthalene-1,2-dioxygenase; [AF009224], [PSEBEDC12A] benzoate-1,2-dioxygenase; [PWWXYL] toluate dioxygenase; [ASCBAABC], [U18133] 3-chlorobenzoate-3,4-dioxygenase; [PCCBDABC] 2-chlorobenzoate-1,2-dioxygenase; [BSU62430] 2,4-dinitrotoluene dioxygenase; [PSU49504] 2-nitrotoluene dioxygenase; [PPU24215] p-cumate-2,3-dioxygenase; [U18133] 3-chlorobenzoate 3,4-(4,5)-dioxygenase; [PSEPHT] phthalate-4,5-dioxygenase; [AB008831], [ACCANI], [D85415] aniline 1,2-dioxygenase; [D90884] phenylpropionic acid 2,3-dioxygenase; [PPPOBAB] phenoxybenzoate dioxygenase; [AF060489], [AB001723], [D89064]carbazole dioxygenase.

Also of utility are organisms whose genomes contain genes encoding other dioxygenases, including tetralin-5,6-dioxygenase, Sikkema *et al.*, *Appl. Eviron. Microbiol.* **59**:567-573, (1993); p-cumate-2,3-dioxygenase DeFrank *et al.*, *J. Bacteriol.* **129**:1356-1364 (1977); fluorenone 1,1a-dioxygenase, Selifonov *et al*, *Biochem. Biophys. Res. Comm.* **193**:67-76(1993); dibenzofuran-4,4a dioxygenase, Trenz *et al*, *J. Bacteriol.* **176**:789-795 (1994); phthalate-3,4-dioxygenase, Eaton *et al.*, *J. Bacteriol.* **151**:48-58 (1982); and 2-chlorobenzoate-1,2-dioxygenase (Selifonov *et al*, *Biochem Biophys. Res. Comm.* **213(3)**:759-767 (1995).

Solely for the purpose of illustrating this invention with respect to oxidation of non-aromatic π-bonds, nucleic acids encoding arene dioxygenases, or any fragments

thereof, are preferably selected from the following non-limiting set of genes and organisms: naphthalene 1,2-dioxygenase, 2,4-dinitro-toluene-4,5-dioxygenase, 2-nitrotoluene 2,3-dioxygenase, toluene-2,3-dioxygenase, isopropylbenzene-2,3-dioxygenase, benzene-1,2-dioxygenase and biphenyl-2,3-dioxygenase, chlorobenzene and tetrachlorobenzene dioxygenases.

Additional suitable homologous arene dioxygenase genes can be found in many microorganisms which one skilled in the art can isolate from various sources including, for example, soil, sediment, air, and aqueous samples by enrichment culture techniques in mineral media using aromatic compounds such as alkyl and halogen-substituted benzenes, biphenyls, indans, naphthalenes and tetralins as carbon sources.

b. Oxidation of aromatic π-bonds

The present invention also provides improved polypeptides capable of oxidizing aromatic π-bonds for producing oxygen-containing aromatic species. Although substantially any oxidized aromatic or heteroaromatic species can be obtained using the polypeptides of the invention, in a presently preferred embodiment, the species include hydroxylated aromatic carboxylic acids and hydroxy alkyl arenes.

Hydroxylated aromatic compounds, such as hydroxylated aromatic carboxylic acids, and alkyl hydroxyarenes (*e.g.*, di- and tri-methyl phenols) are an important group of industrial chemicals. For example, the methylphenols find use in the industrial synthesis of vitamin E, and, also in the synthesis of various polymers and resins, where they can be used individually or as part of more complex compositions that include other phenolic and non-phenolic compounds. Typically, dimethylphenols and trimethylphenols are generated by successive methylation of phenol and cresols. However, current synthetic chemical methods based on methylation of phenol do not offer sufficiently high selectivity for preparing isomeric dimethyl- and trimethyl-phenols individually. Therefore, complex mixtures of methylated products having a varying degree of methylation and various substitution patterns are generally obtained. The presence of these mixtures necessitates the use of expensive separation procedures to obtain pure compounds.

Hydroxylated aromatic carboxylic acids (HCA) and their esters and lactones are useful components of various polymers and co-polymers, such as polyesters. Their utility stems largely from their bifunctional reactive nature (*e.g.*, hydroxyl and carboxyl groups) and the hydrophobic aromatic ring, which often imparts desirable physical and

chemical properties to the polymers (p-hydroxybenzoate, m-hydroxybenzoate). HCAs are also used in, for example, anti-microbial additives to pharmaceuticals (esters of para-hydroxybenzoic acid, parabens) and to fragrances (esters of salicylic acid, coumarins and 3,4-dihydrocoumarin).

While many chemical synthetic methods for HCAs and their derivatives are known in the art, typically, these compounds are manufactured from a non-oxidized aromatic precursor by multistep processes, requiring both the harsh conditions and the extensive product purification from by-products, arising from non-selective reactions.

The present invention solves many of these problems by providing a dioxygenase polypeptide that oxidizes one or more aromatic π-bonds to the corresponding diol. The diols can, if desired, be subsequently dehydrated to restore the aromatic system and yield a hydroxylated aromatic ring.

C. Accessory Polypeptides

In conjunction with the oxidative pathways utilizing polypeptides having dioxygenase activity, as discussed above, the present invention provides accessory non-dioxygenase polypeptides. As used herein, "accessory polypeptides" refers to those polypeptide that do not carry out the initial dioxidation step in the methods of the invention. Exemplary accessory polypeptide include, ligases, transferases, dehydrogenases, and the like. Although both shuffled and non-shuffled accessory polypeptides can be used, preferred accessory polypeptides are those that have been shuffled.

The non-dioxygenase polypeptides can be used at any step of a pathway using a dioxygenase of the invention. In a preferred embodiment, the accessory polypeptides are used to further transform an oxidation product. Although it will generally be preferred to utilize oxidized substrates that are produced by a dioxygenase of the invention, those of skill will appreciate that these routes can be practiced with analogous substrates that are, for example chemically synthesized, commercially available, *etc.*

Moreover, the present invention provides methods using both the improved accessory peptides and unimproved accessory peptides to further elaborate the dioxygenase-mediated reaction product. The method involves contacting the product of the dioxygenase-mediated reaction with one or more of the accessory polypeptides. In a preferred embodiment, the product is contacted with an organism that expresses the accessory

polypeptide(s). When the accessory polypeptides are improved polypeptides, they will generally be produced by the methods described herein.

The improved dioxygenase and the accessory polypeptide(s) can be expressed by the same host cell, or they can be expressed by different host cells. In a preferred embodiment, the accessory polypeptide and the improved dioxygenase are expressed by the same host cell.

By utilizing accessory polypeptides, the present invention makes possible the synthesis of a great variety of industrially valuable compounds via the methods disclosed herein.

1. Dehydrogenases

In a preferred embodiment, an alcohol or diol is converted to an aldehyde or carboxylic acid by the action of a dehydrogenase. The substrate for the dehydrogenase is preferably the product of an improved oxygenase of the invention.

Polynucleotides encoding many known dehydrogenases can be used as substrates for nucleic acid shuffling. Exemplary dehydrogenases useful in practicing the present invention include, but are not limited to:

[ECOALDB, ECAE000436, ECAE000239, D90780, D90781, ECOFUCO, ECOFUCO] dehydrogenase of *Escherichia coli*; [AF029734 and AF029733] dehydrogenase of *Xanthobacter autotrophicus*; [AREXOYGEN] dehydrogenase of *Agrobacterium radiobacter*; [AB003475] dehydrogenase of *Deinococcus radiodurans*; [AF034434, VIBTAGALDA] dehydrogenase of *Vibrio cholerae*; [D32049] dehydrogenase of *Synechococcus* sp.; [AE001154] dehydrogenase of *Borrelia burgdorferi* (BB0528); [ABY17825] dehydrogenase of *Agaricus bisporus*; [ASNALDAA] dehydrogenase of *Aspergillus niger*; [EMEALDA, EMEALCA] dehydrogenase of *Aspergillus nidulans;* [AF019635, PPU15151] dehydrogenase of *Pseudomonas putida* TOL plasmid, xylW, xyl C; [AF031161] dehydrogenase of *Pseudomonas sp.* VLB120, (stdD); [PFSTYABCD] dehydrogenase of *P. fluorescens*, styD; [PPU24215] dehydrogenase of *P. putida*, F1*p*-cymene alcohol and aldehyde dehydrogenases.

2. *Conversion of hydroxyls and/or acids to esters*

In another preferred embodiment, a method is provided for converting carboxylic acid and hydroxyl groups to adducts such as esters and ethers. Useful polypeptides include, for example, ligases and transferases (*see*, **Fig.13**). For the purposes of the discussion below, these polypeptides are referred to as "adduct-forming" polypeptides.

The adduct-forming polypeptides are useful for enhancing and controlling the production of biotransformation products. These polypeptides, which convert a diol, for example, to a monoacyl or monoglycosyl derivative can enhance control over the regioselectivity of subsequent reactions (*e.g.*, chemical dehydration). For example, the regioselectivity of chemical dehydration in certain cases can be controlled by converting the compounds to their diacyl derivatives by means of chemical reaction, and then selectively removing one of the acyl groups using an polypeptide of the invention. Alternatively, one can control the regioselectivity of the dehydration by using an esterase or a trans-acylase polypeptide to convert the compounds to monoacyl derivatives, preferably in the presence of an excess of another carboxylic acid ester. In addition, the isolation of certain products is simplified by their conversion to more hydrophobic species. For example, the acylation of a diol to the corresponding carboxylic ester provides a more efficient recovery of such diols, in the form of an ester, by organic solvent extraction of the adduct. Preferred organic solvents are those that can be used in an immiscible biphasic organic-aqueous biotransformation with whole cells, whether in a batch or in a continuous mode.

An adduct-forming polypeptide is optionally expressed by the same host cell that expresses the dioxygenase, dehydrogenase, racemase, *etc.*, or by a different host cell. Moreover, an adduct-forming polypeptide can be a naturally occurring polypeptide, or it can be improved by the method of the invention.

When the adduct-forming polypeptide is an improved polypeptide, in presently preferred embodiments, the polypeptide demonstrates increased efficiency in the formation of the monoacyl- or monoglycosyl- derivatives of a desired compound (*e.g.*, a glycol, carboxylic acid, *etc.*). Other improved adduct-forming polypeptides include transferases and ligases that selectively modify only one of the hydroxyl groups of a diol, thus providing a means for controlling the regioselectivity of dehydration of such derivatives to either of two possible isomeric α-hydroxycarboxylic acid compounds.

a. Acyltransferases

Other enzymes useful in practicing the present invention are the acyltransferases. These polypeptides are optionally evolved to enhance certain catalytic properties of the encoded polypeptides such as, specificity for a particular hydroxyl and/or acid, enantiomeric and/or diastereomeric selectivity.

More specifically, these polypeptides catalyze acyl transfer reactions as shown in **Fig. 13**. Acyltransferases are ubiquitous in nature, and many organisms (*e.g.*, microbes, plants, mammals, *etc.*) can be used as sources of genes encoding these polypeptides. No matter their origin, the acyltransferase genes are preferably selected from those encoding functional polypeptides that catalyze active (CoA) ester transfer reactions in the biocatalytic processes described herein. Preferred acyltransferase genes are selected from those encoding functional polypeptides catalyzing reactions of small non-biopolymeric molecules.

Examples of various acyltransferases useful in the present invention include polypeptides that catalyze the methylation of α-hydroxycarboxylic acids. A list of exemplary polynucleotides that can be recruited for this purpose are listed below by the corresponding GenBank identification:

[AF043464] acetyl-CoA: benzylalcohol acetyltransferase of *Clarkia breweri*, and benzoyl-CoA benzyl alcohol acetyltransferase present in the same organism, (Dudareva *et al, Plant Physiol.* **116(2)**:599-604 (1998)); [DCANTHRAN, DCHCBT1, DCHCBT1A, DCHCBT1B, DCHCBT2, DCHCBT3] hydroxycinnamoyl/benzoyl-CoA:anthranilate N-acyltransferase of *Dianthus caryophyllus*; [E08840] homoserine o-acetyltransferase of *Acremonium chrysogenum*; [E12754] anthocyanin 5-aromatic acyltransferase, of *Gentiana triflora*; [HUMBCAT] branched chain acyltransferase (human, J03208, J04723); [MG396;D02 °orf152(lacA); MJ1064(lacA) MJ1678, MTH1067]; galactoside 6-O acetyl transferase EC 2.3.1.18, lac A of *E.coli* ; B0342(lacA); or of other organisms; [B3607(cysE), HI0606(cysE), HP1210(cysE), SLR1348(cysE)] serine O-acetyltransferase EC 2.3.1.30; [YGR177C, YOR377W] alcohol O-acetyltransferase, EC 2.3.1.84, of *Saccharomyces cerevisiae*; [*e.g.*, Q00267,D90786,Z92774,I78931 AF030398, AF008204, AF042740] arylamine N-acetyltransferase, EC 2.3.1.118; [YAR035(YAT1),

YM8054.01(CAT2)] carnitine O-acetyltransferase, EC 2.3.1.7, or mammalian origin of from yeast; [CHAT] choline O-acetyltransferase, EC 2.3.1.6, of mammalian origin; acetyl CoA:deacetylvindoline 4-O-acetyltransferase (EC 2.3.1.107) St-Pierre *et al, Plant J.* 14(6): 703-713 (1998); and [ECOPLSC] 1-acyl-sn-glycerol-3-phosphate acyltransferase (plsC) of *Escherichia coli.*

b. Acyl CoA ligases

In another embodiment an accessory polypeptide having acyl CoA ligase activity is provided.

Specificity of acyl-CoA ligases towards a particular exogenous substrate or a group of substrates is preferably optimized by screening or selecting for the acylation of a substrate by shuffled and co-expressed acyl-CoA ligases and acyltransferases. Utilizing these polypeptides in tandem allows the combined effect of both polypeptides to be exploited.

To illustrate the single gene shuffling or shuffling of a family of genes approach to improve acyl-CoA ligases or acyltransferases, one or more of the members of the corresponding superfamilies of these polypeptides are selected, aligned with similar homologous sequences, and shuffled against these homologous sequences.

An exemplary list of useful acyl-CoA ligase genes for inclusion into an organism of the invention is provided below:

[AF029714, ECPAA, AJ000330, PSSTYCATA] phenylacetate-CoA ligase, EC 6.2.1.30; [Y11070, Y11071] phenylpropionate-CoA ligase; [B2260(menE), SLR0492(menE), SAU51132(menE)] O-succinylbenzoate-CoA ligase, EC 6.2.1.26; [RPU75363, RBLBADA, AA532705, AA664442, AA497001, AF042490, ARGFCBABC] (chloro)benzoate-CoA ligase, EC 6.2.1.25; [SBU23787, VPRNACOAL, POTST4C11, RIC4CL2R, OS4CL, AF041051, AF041052, GM4CL14, GM4CL16, LEP4CCOALA, LEP4CCOALB, PC4CL1A, PC4CL1AA, PC4CL2A, PC4CL2AA, TOB4CCAL, TOBTCL2, TOBTCL6, ECO110K, AF008183, AF008184, AF041049, AF041050, ATU18675, NTU5084, NTU50846, PTU12013, PTU39404, PTU39405, ATF13C5, ORU61383, AF064095, AA660600, AA660679, STMPABA] 4-coumarate-CoA ligase EC 6.2.1.12; [RPU02033] 4-hydroxybenzoate-CoA ligase; [PSPPLAS] 2-aminobenzoate-CoA ligase.

In some embodiments of the invention, a carboxylic acid is fed exogenously to an organism that expresses the ligase or transferase. Preferably, the carboxylic acid is selected from those compounds that cannot be altered by the polypeptide used to produce the substrate acted upon by the adduct forming polypeptide. Such carboxylic acids include, for example, both substituted and non-substituted benzoic acid, phenylacetic acid, naphthoic, phenylpropionic acid, phenoxyacetic acid, cycloalkanoic acid, carboxylic acids derived from terpenes, pivalic acid, substituted acrylic acids, and the like.

To facilitate the utilization of exogenously supplied carboxylic acids, and for enhancing the variety of compounds suitable for use in this process, the invention also provides microorganisms in which one or more mutations are introduced. Preferred mutations are those that effectively block metabolic modifications of such acids beyond their conversion to a suitable active ester (*e.g.*, as a derivative of coenzyme A). Such mutations in the host organism are optionally introduced by classical mutagenesis methods, by site-directed mutagenesis, by whole genome shuffling, and other methods known to those of skill in the art. One can also introduce mutations that minimize host endogenous esterase activity.

In a presently preferred embodiment, the acyl transferase-encoding nucleic acids used as substrates for creating recombinant libraries encode polypeptides that transfer an acetyl group from an endogenous pool of acetyl-CoA in the cells of the host. The endogenous pools of acetyl-CoA can also be enhanced by nucleic acid shuffling of an acetyl-CoA ligase and by supplying an exogenous acetate in the medium.

While using acetyl-CoA transferases or other acyltransferase or glycosyltransferase does not necessarily require expression of a corresponding acetyl-CoA or other ligase, in a presently preferred embodiment, the organisms produce a sufficient amount of an acyl-CoA ligase so as to activate the carboxylic acids to CoA thioesters, which in turn serve as substrates for acyl-CoA transferases that utilize the oxidation products as substrates. The specificity of an acyl-CoA ligase towards a desired exogenous carboxylic acid can be optimized using the recombination and screening/selection methods of the invention. Preferably, the screening or selecting is performed using co-expressed acyl-CoA ligases and acyltransferases, thus permitting one to screen on the basis of the combined effect of both polypeptides in the pathway for provision of monoacylated derivatives of the oxidation products.

Nucleic acids that encode acyl-CoA ligases and other acyltransferases useful as substrates for the recombination and selection/screening methods of the invention include, for example, one or more members of the superfamilies of these polypeptides. In a presently preferred embodiment, the nucleic acids are selected, aligned with similar homologous sequences, and shuffled against these homologous sequences.

c. Glycosyltransferases

Similarly, one or more glycosyltransferases can be expressed by the host cells of the invention. Alternatively, one or more glycosyltransferases can be selected from the glycosyltransferase superfamily, aligned with similar homologous sequences, and shuffled against these homologous sequences. Glycosyl transfer reactions are ubiquitous in nature, and one of skill in the art can isolate such genes from a variety of organisms, using one or more of several art-recognized methods. The following are illustrative examples of glycosyltransferase-encoding nucleic acids that can be used as substrates for creation of the recombinant libraries. The libraries are then screened to identify those polypeptides that exhibit an improvement in the glycosylation of compounds such as alcohols, diols and α-hydroxycarboxylic acids:

[EC 2.4.1.123] inositol 1-α-galactosyltransferase; [NTU32643, NTU32644] phenol β-glucosyltransferase, EC 2.4.1.35; flavone 7-O-beta-glucosyltransferase, EC 2.4.1.81; [AB002818, ZMMCCBZ1, AF000372, AF028237, AF078079, D85186, ZMMC2BZ1, VVUFGT]; flavonol 3-O-glucosyltransferase, EC 2.4.1.91; o-dihydroxycoumarin 7-O-glucosyltransferase, EC 2.4.1.104; vitexin beta-glucosyltransferase, EC 2.4.1.105; coniferyl-alcohol glucosyltransferase, EC 2.4.1.111; monoterpenol beta-glucosyltransferase, EC 2.4.1.127; arylamine glucosyltransferase, EC 2.4.1.71; sn-glycerol-3-phosphate 1-galactosyltransferase, EC 2.4.1.96; [RNUDPGTR, AA912188, AA932333] glucuronosyltransferase, EC 2.4.1.17; the human UGT and isoenzymes (~35 genes); salicyl-alcohol glucosyltransferase, EC 2.4.1.172; 4-hydroxybenzoate 4-O-beta-D-glucosyltransferase, EC 2.4.1.194; zeatin O-beta-D-glucosyltransferase, EC 2.4.1.203; [VFAUDPGFTA] D-fructose-2-glucosyltransferase; and [MBU41999] ecdysteroid UDP-glucosyltransferase (egt).

In presently preferred embodiments, the glycosyltransferases are selected from those which transfer hexose residues from UDP-hexose derivatives. Preferred hexoses include, for example, D-glucose, D-galactose and D-N-acetylglucosamine.

d. Methyltransferases

In a still further preferred embodiment, the host cells of the present invention express a polypeptide capable of converting a carboxylic acid to a carboxylic acid methyl ester. Presently preferred polypeptides include methyltransferases.

For the purpose of this invention, genes encoding S-adenosylmethionine-dependent methyltransferases are preferred. In a preferred embodiment, these polypeptides are evolved to enhance selected properties of the encoded polypeptides such as, specificity for a particular substrate and enantiomeric and/or diastereomeric selectivity and/or solvent resistance.

More specifically, these polypeptides can be evolved to catalyze the O-methylation of carboxyl groups of a caroxylic acid substrate thus forming the corresponding methyl esters. Methyltransferases are ubiquitous in nature, and many organisms (*e.g.*, microbes, plants, mammals, *etc.*) can be used as sources of genes encoding these polypeptides. No matter their origin, the methyltransferase genes are preferably selected from those which encode functional polypeptides that catalyze the methylation of small non-biopolymeric molecules. Preferably, the methyltransferases are those which act on the carboxyl groups of organic acids.

Examples of various methyltransferases that can be expressed by host cells of the invention and which are useful for nucleic acid shuffling-based directed evolution of polypeptides catalyzing the methylation of carboxylic acids are listed below by the corresponding GenBank identification:

[SCCCAGC3] methyltransferase of *Streptomyces clavuligerus* methyltransferase CmcJ; [SEERYGENE] methyltransferase of *S.erythraea* methyltransferases; [SEU77454] methyltransferase of *Saccharopolyspora erythraea*; erythromycin O-methyltransferase (eryG); [SGY08763] methyltransferase of *S.griseus*; [SKZ86111] methyltransferase of *S.lividans*; [STMDNRDKP] methyltransferase of *Streptomyces peucetius*; carminomycin o-methyltransferase (dnrK); [MDAJ39670] methyltransferase of *Streptomyces ambofaciens*; [SEY14332] methyltransferase of

Saccharopolyspora erythraea; [SPU10405] methyltransferase of *Streptomyces purpurascens* ATCC 25489; [STMDAUA] methyltransferase of *Streptomyces* sp.; aklanonic acid methyltransferase (dauC), and carminomycin 4-O-methyltransferase (dauK); [SC2A11 and SC3F7] methyltransferase of *Streptomyces coelicolor*; [SHGCPIR] methyltransferase of *S.hygroscopicus;* [STMCARMETH] methyltransferase of *Streptomyces peucetius* carminomycin 4-O-methyltransferase; [STMODPOMT] methyltransferase of *Streptomyces alboniger* O-demethylpuromycin-O-methyltransferase (dmpM); [STMTCREP]; methyltransferase of *Streptomyces glaucescens*; [SLLMRBG] methyltransferase of *S. lincolnensis* lmrB methyltransferase; [SSU65940] 31-O-demethyl-FK506 methyltransferase (fkbM) of *Streptomyces* sp.; [STMDAUABCE] aklanonic acid methyltransferase (dauC) of *Streptomyces* sp.; [STMMDMBC] O-methyltransferase (mdmC) of *Streptomyces mycarofaciens*; [STMTYLF] macrocyn-O-methyltransferase (tylF) of *S.fradiae*; [E08176] Gene of mycinamicin III-O-methyltransferase; [AF040571] methyltransferase of *Amycolatopsis mediterranei*; [ECU56082] S-adenosylmethionine:2-demethylmenaquinone methyltransferase (menG) of *Escherichia coli*; [RHANODABC] methyltransferase (nodS) of *Azorhizobium caulinodans*; [YSCSTE14] isoprenylcysteine carboxyl methyltransferase (STE14) of *Saccharomyces cerevisiae*; [YSCMTSW] farnesyl cysteinecarboxyl-methyltransferase (STE14) of *Saccharomyces cerevisiae*; [YSCDHHBMET] 3,4-dihydroxy-5-hexaprenylbenzoate methyltransferase (COQ3) of *S.cerevisiae*; [AF004112 and AF004113] phospholipid methyltransferases (cho1+), (cho2+) of *Schizosaccharomyces pombe*; [ASNOMT, ASNOMT1A, ASNOMT1B, ASNOMT1C and AF036808-AF036830] O-methyltransferases of *Aspergillus*; [MSU20736] S-adenosyl-L-methionine; trans-caffeoyl-CoA3-O-methyltransferase of *Medicago sativa*; [ALFIOM] isoliquiritigenin 2'-O-methyltransferase of *Medicago sativa*; [MSU20736] S-adenosyl-L-methionine; trans-caffeoyl-CoA3-O-methyltransferase (CCOMT) of *Medicago sativa*; [MSAF000975] 7-O-methyltransferase (7-IOMT(6)) of *Medicago sativa*; [MSAF000976] 7-O-methyltransferase (7-IOMT(9)) of *Medicago sativa*; [MSU97125] of isoflavone-O-

methytransferase *Medicago sativa*; [NTCCOAOMT] caffeoyl-CoA O-methyltransferase of *Nicotiniana tabacum*; [NTZ82982] caffeoyl-CoA O-methyltransferase 5 of *N.tabacum*; [NTDIMET] o-diphenol-O-methyltransferase of *N.tabacum*; [PCCCOAMTR, PUMCCOAMT] trans-caffeoyl-CoA 3-O-methyltransferase of *Petroselinum crispum*; [PTOMT1] *s* caffeic acid/5-hydroxyferulic acid O-methyltransferase (PTOMT1) of *Populus tremuloide*; [PBTAJ4894-PBTAJ4896] caffeoyl-CoA 3-O-methyltransferases of *Populus balsamifera* subsp. trichocarpa; [ZEU19911] S-adenosyl-L-methionine: caffeic acid 3-O-methyltransferase of *Zinnia elegans*; [SLASADEN] S-adenosyl-L-methionine:trans-caffeoyl-CoA 3-O-methyltransferase of *Stellaria longipes*; [VVCCOAOMT] caffeoyl-CoA O-methyltransferase of *V.vinifera*; [D88742] O-methyltransferase of *Glycyrrhiza echinata*; [AF046122] caffeoyl-CoA 3-O-methyltransferase (CCOMT) of *Eucalyptus globulus*; [ATCOQ3] dihydroxypolyprenylbenzoate: methyltransferase of *Arabidopsis thaliana* [CSJSALMS9O] S-adenosyl-L-methionine:scoulerine 9-O-methyltransferase of *Coptis japonica*; [HVU54767] caffeic acid O-methyltransferase (HvCOMT) of *Hordeum vulgare*; [MCU63634] inositol methyltransferase (Imt1) of *Mesembryanthemum crystallinum*; [PSU69554] 6a-hydroxymaackiain methyltransferase (hmm6) of *Pisum sativum*; [CAU83789] O-diphenol-O-methyltransferase of *Capsicum annuum*; [U16794] 3' flavonoid O-methyltransferase (fomt1) of *Chrysosplenium americanum*; [CBU86760] SAM:(Iso)eugenol O-methyltransferase(IEMT1) of *Clarkia breweri*; salicylic acid carboxyl SAM-O-methyltransferase (Dudareva *et al, Plant Physiol.* **116**(2):599-604 (1998)); [HSHIOMT9] hydroxyindole-O-methyltransferase (HIOMT) of *Homo sapiens*; [HSCOMT2] gene catechol O-methyltransferase of *Homo sapiens*; [HUMPNMTA] phenylethanolamine N-methyltransferase gene of *Homo sapiens*; [HUMCOMTA] catechol-O-methyltransferase of *Homo sapiens*; [HUMCOMTC] catechol-O-methyltransferase of *Homo sapiens*; [HUMPNMT] phenylethanolamine N-methyltransferase of *Homo sapiens*; [AF064084] prenylcysteine carboxyl methyltransferase (PCCMT) of *Homo sapiens*; [HUMCMT] carboxyl methyltransferase of *Homo sapiens*;

[HUMHNMA] histamine N-methyltransferase of *Homo sapiens*; [RATCATAA, RATCATAB] catechol-O-methyltransferase of *R.norvegicus*; [RATDHNPBMT] dihydroxypolyprenylbenzoate methyltransferase of *Rattus norvegicus*; [BOVPNMTB] of Bovine phenylethanolamine N-methyltransferase; [MPEMT7] phosphatidylethanolamine-N-methyltransferase of *Mus musculus* 2; [MMU86108] nicotinamide N-methyltransferase (NNMT) of *Mus musculus*; [MUSCMT] carboxyl methyltransferasease protein of Mouse; [GDHOMT] hydroxyindole-O-methyltransferase of *G.domesticus*; [DRU37434] L-isoaspartate (D-aspartate) O-methyltransferase (PCMT)of *Danio rerio*; [DMU37432] protein D-aspartyl, L-isoaspartylmethyltransferase of *Drosophila melanogaster*; and [HAU25845 and HAU25846] farnesoic acid o-methyl-transferases of *Homarus americanus*.

3. Enantiomeric interconversion.

In a still further preferred embodiment, the present invention provides a nucleic acid encoding a polypeptide capable of converting a particular enantiomer of a chiral compound such as an alcohol, diol or α-hydroxycarboxylic acid or a precursor or analogue thereof to its antipode.

Presently preferred polypeptides include racemases, such as the mandelate racemase of *Pseudomonas putida* (PSEMDLABC). These polypeptides can expressed by hosts of the invention in their natural form or, alternatively, they can be evolved to enhance certain catalytic properties of the encoded polypeptides such as, specificity for a particular substrate and enantiomeric and/or diastereomeric selectivity.

The nucleic acids encoding the mandelate racemase of *Pseudomonas putida*, which catalyzes the interconversion of mandelate R and S enantiomers, is a typical preferred example of genes selected for use in this invention. The nucleic acids encoding this gene, and any homologs of thereof, are subjected to nucleic acid shuffling to evolve polypeptides having improved or optimal performance and specificity towards particular substrates such as α-hydroxycarboxylic acids. In a preferred embodiment, the polypeptide has a performance and/or specificity that is enhanced over the wild type. Preferred polypeptides act on α-hydroxycarboxylic acid substrates, such as those displayed in **Fig. 11.**

4. Solvent resistance polypeptides

The invention also provides organisms expressing one or more of the improved polypeptides of the invention and that are also resistant to solvents, organic substrates and reaction products (*e.g.*, epoxides, glycols, α-hydroxyaldehydes, α-hydroxycarboxylic acids and α-hydroxycarboxylic acid derivatives (*e.g.*, esters)) according to the methods of the invention.

The solvent resistance of organisms and polypeptide used in the biocatalytic conversion of organic compounds is important for enhancing the productivity of such processes. Increased solvent resistance of the organisms can enhance longevity, viability and catalytic activity of the microbial cells, and can simplify the administration of the feedstock compounds to the reactor and the recovery or separation of desired products by means of, for example, continuous or semi-continuous liquid-liquid extraction.

In another aspect, the invention provides microbial cells that are useful in the synthetic methods described herein, which express proteins conferring resistance to solvents (in particular, organic solvents) upon the microbial cells. This allows the use of whole microbial cells in a organic-aqueous mixture (*e.g.*, a biphasic mixture). In presently preferred embodiments, the invention provides microbial strains including at least two of the polypeptide systems described herein. For example, a microorganism of the invention can contain both a dioxygenase gene and a transferase gene. In other embodiments, the microorganism can contain both an arene dioxygenase gene and a solvent resistance gene. The microbial cells thus provide a significant improvement in productivity of the synthesis processes, selectivity of product formation, operational simplicity, ease of product recovery and minimizing any by-product streams.

Several microorganisms are known to possess high resistance to hydrophobic compounds such as benzene and lower alkylbenzenes. Recently, genes encoding a solvent efflux pump (*srp*ABC) have been identified in *Pseudomonas putida* strains (Kieboom *et al.*, *J. Biol. Chem.* **273**:85-91 (1998)). Similarly, various genes that encode polypeptides that confer organic solvent resistance can be found in bacterial strains such as *Pseudomonas putida* GM73 (Kim *et al.*, *J. Bacteriol.* **180**: 3692-3696 (1998)), *Pseudomonas putida* DOT-T1E (Ramos *et al.*, *J. Bacteriol.* **180**: 3323-3329 (1998)), *Pseudomonas idaho* (Pinkart and White, *J. Bacteriol.* **179**: 4219-4226 (1997)). These and other genes, such as those that encode many proton-dependent multidrug efflux systems, *e.g.*, MexA-MexB-OprM, MexC-

MexD-OprJ, and MexE-MexF-OprN of *Pseudomonas aeruginosa* (Li *et al., J. Bacteriol.* **180**: 2987-2991 (1998)), or the *tolC, acrAB, marA, soxS,* and *robA* loci of *Escherichia coli* (Aono *et al., J. Bacteriol.* **180**: 938-944 (1998); White *et al., J. Bacteriol.* **179**: 6122-6126 (1997)), and in many other microorganisms, can be used to confer solvent resistance upon a host microbial strain used in the oxidative biocatalytic conversion of olefins by action of dioxygenases.

In presently preferred embodiments, the ability of a polypeptide to confer solvent resistance is enhanced by subjecting nucleic acids encoding solvent resistance polypeptides, or the genomes of the microorganisms themselves, to the recombination and selection/screening methods described herein. The nucleic acids listed above, as well as similar genes, provide a source of substrates for incorporation into organisms of the invention and/or use in nucleic acid shuffling and other methods of constructing libraries of recombinant polynucleotides. The libraries can then be screened to identify those nucleic acids that encode polypeptides conferring improved solvent tolerance on a host. For example, one can select for improved tolerance to compounds such as olefins, AHAs, aldehydes, esters and hydrophobic solvents, including alkanes, cycloalkanes, alcohols and halocarbon derivatives, for example, which are used for performing biotransformation (*e.g.*, two-phase oxidation) of olefins to glycols, AHAs and to their corresponding acyl- and glycosyl- derivatives, *etc.* Similarly, shuffling of nucleic acids that encode these polypeptides can be used to confer and to improve resistance of the microbial cell to high concentrations of biotransformation substrates, intermediates and endproducts, thus improving biocatalyst performance and productivity.

In addition to each of the methods set forth above, the present invention provides polypeptides produced according to these disclosed methods. Moreover, the invention provides organisms that express the polypeptides produced by the method of the invention. The organisms of the invention can express one or more of the improved polypeptides. Also provided by the present invention are methods of synthesizing a desired compound. This method involves contacting an appropriate substrate with a polypeptide of the invention. In a preferred embodiment, the substrate is contacted with an organism of the invention that expresses a polypeptide of the invention.

D. Methods of Using Improved Polypeptides to Prepare Organic Compounds

In addition to the methods discussed above, the present invention provides a range of methods for preparing useful organic compounds by the oxidation and further elaboration of appropriate precursors. Among the methods provided by the present invention are, for example, the oxidation of alkylarene compounds to the corresponding unsaturated diols and the subsequent dehydration of these diols to hydroxy alkylarenes. Additionally, there is provided an analogous method for preparing hydroxylated aromatic carboxylic acids. Moreover, the invention provides methods for preparing exocyclic and/or acyclic diols from molecules having alkene bonds. These diols can be readily converted to α-hydroxycarboxylic acids.

The reaction types and sequences set forth below are illustrative of the scope of the invention. The dioxygenases of the invention are capable of oxidizing any organic substrate comprising an oxidizable moiety. Additional reaction sequences utilizing the polypeptides of the invention will be apparent to those of skill in the art.

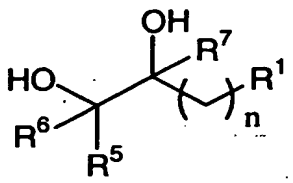
1. Preparation of vicinal diols

The formation of vicinal diols by oxidizing a π-bond using a dioxygenase of the invention provides ready access to a wide array of compounds that are useful as both final products and as intermediates in multi-step reaction pathways. The dioxygenases of the invention are capable of converting to vicinal diols an array of structurally distinct compounds comprising one or more π-bonds.

Although the method can be practiced with essentially any π-bond, in essentially any compound, in a preferred embodiment, the method involves preparing a vicinal diol group by contacting a substrate comprising a carbon-carbon double bond with an improved dioxygenase polypeptide, or an organism expressing an improved dioxygenase polypeptide.

In another preferred embodiment, the substrate comprising the carbon-carbon π-bond is selected from styrene, substituted styrene, divinylbenzene, substituted divinylbenzene, isoprene, butadiene, diallyl ether, allyl phenyl ether, substituted allyl phenyl ether, allyl alkyl ether, allyl aralkyl ether, vinylcyclohexene, vinylnorbornene, and acrolein.

In yet another preferred embodiment, the vicinal diol produced by the action of the improved dioxygenase polypeptide has the structure:



wherein R^1 and R^5 are each independently selected from alkyl, substituted alkyl, aryl, substituted aryl, heteroaryl, substituted heteroaryl, heterocyclic, substituted heterocyclic, —NR^2R^3, —OR^2, —CN, $C(R^4)NR^2R^3$ and $C(R^4)OR^2$ groups, or R^1 and R^5 are joined to form a ring system selected from saturated hydrocarbon rings, unsaturated hydrocarbon rings, optionally substituted, saturated or unsaturated heterocyclic rings; R^2 and R^3 are members independently selected from H, alkyl, substituted alkyl, aryl, substituted aryl, heteroaryl, substituted heteroaryl, heterocyclic and substituted heterocyclic groups; R^4 is selected from =O and =S; R^6 and R^7 are independently selected from H and alkyl; and n is an integer from 0 to 10, inclusive.

In certain preferred vicinal diols, R^1 is selected from phenyl, substituted phenyl, pyridyl, substituted pyridyl, —NR^2R^3, —OR^2, —CN, $C(R^4)NR^2R^3$ and $C(R^4)OR^2$ groups, R^2 and R^3 are members independently selected from H, alkyl, substituted alkyl, aryl, substituted aryl, heteroaryl, substituted heteroaryl, heterocyclic and substituted heterocyclic groups; and R^4 is selected from =O and =S.

In another preferred embodiment, the diol includes a six-member ring having at least one endocyclic double bond and at least one substituent selected from methyl, carboxyl and combinations thereof. Preferred diols having this structure are displayed in **Fig.1** and are the compounds having the structures III, IV, V, VI, VII, VIII, and **Fig. 4** and are compounds having the structures XXIII, XXIV, XXV, XXVI, XXVII, XXVIII, XXIX, XXX.

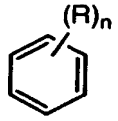
2. Preparation of hydroxy alkylarenes

In another preferred embodiment, the invention provides methods for preparing hydroxy arenes and hydroxy arenes that are further functionalized with, for example, alkyl and substituted alkyl groups. The method involves contacting a substrate comprising an aryl group with an improved dioxygenase of the invention to from a diol. The diol intermediate is subsequently dehydrated, thereby producing an aromatic ring functionalized with a hydroxy radical. Methods for carrying out the dehydration are known

in the art. Both enzymatic and chemical/physical means are appropriate. Preferred chemical/physical techniques include acid, base, heat and combinations thereof.

In another preferred embodiment, the substrate includes a member selected from arylalkyl groups, substituted arylalkyl groups, heteroarylalkyl groups, and substituted heteroarylalkyl groups.

In a still further preferred embodiment, the substrate has the structure



wherein, each of the *n* R groups is a member selected from the group consisting of alkyl groups, substituted alkyl groups, alkynyl groups, aralkyl groups, alkoxy groups, aryloxy, alkylthio groups, cycloalkyl groups, alkenyl groups, halogens, CF_3, CN, NO_2, trimethylsilyl, trimethylgermanyl, trimethylstannyl, and alkylamines; and n is an integer from 0 to 5, inclusive. Typically R contains about 1 to about 15 carbons, preferably R is a lower alkyl group, more preferably methyl; and *n* is an integer from 1 to 5, inclusive, more preferably, *n* is an integer from 1 to 4, inclusive.

The enzymes, bioengineered pathways, and microorganisms of the invention are useful for the synthesis of a wide variety of compounds, including many that are of commercial importance for purposes such as vitamin production. The methods and enzymes of the invention provide a means by which commercially valuable compounds can be formed using relatively inexpensive compounds as precursors. An illustrative example of the use of the enzymes and methods of the invention is a new selective process for making isomeric trimethylphenols which can further be converted to trimethylhydroquinone (a key vitamin E intermediate). In the case of 2,4,5-trimethylphenol, the new route uses similar conditions as are typically applied to conversion of 2,4,6-trimethylphenol (**Fig. 2**).

The use of arene dioxygenases that are optimized using the methods of the invention to synthesize precursors of vitamin E, for example, provide significant advantages over previously available methods for obtaining vitamin E. Natural tocopherols are isolated from common oils such as soybean, corn, canola, cottonseed, safflower, and the like, by means of repeated vacuum distillation and alkali treatment of the oils, followed by multi-step processes for removal of impurities such as sterols and fatty acids. Microbial production of tocopherols has also been reported, from *e.g.*, *Aspergillus*, *Lactobacter*, *Euglene* and *Mycobacterium*, but fermentation titers are low, rendering these processes

commercially insignificant. Chemical synthesis of vitamin E from petrochemicals, for example, typically results in an equal mixture of eight different isomers, with the mixture containing 12.5 percent of each isomer. The recombination and selection methods of the invention can be used to obtain arene dioxygenases that produce predominantly the desired RRR-α-tocopherol by virtue of the enhanced enantiospecificity and regiospecificity of the improved recombinant enzymes.

Key intermediates in the synthesis of D-α-tocopherol (vitamin E), for example, are 2,3,5-trimethylphenol (Compound IX in **Fig. 2**), 2,4,5-trimethylphenol (Compound X), 2,3,6-trimethylphenol (Compound XI) and 2,4,6-trimethylphenol (Compound XII). From these trimethylphenol compounds, one can use catalytic oxidation to obtain 2,3,5-trimethylhydroquinone as shown in **Fig. 2**. One method that is suitable for the catalytic oxidation is described in, for example, US Patent No. 4,250,335. The 2,3,5-trimethylhydroquinone, in turn can be used to prepare vitamin E by any of several known methods. For example, a natural or synthetic phytol side chain can be attached to the 2,3,5-trimethylhydroquinone using Friedel-Craft acidic catalytic conditions (*see, e.g.*, US Patent No. 5,468,883).

The 2,3,5-, 2,4,5, 2,4,6- and 2,3,6-trimethylphenols can be synthesized using the arene dioxygenases and microbial cells of the invention as shown in **Fig. 1**. For example, 1,2,4-trimethylbenzene (Compound I) is oxidized to any of Compounds III, IV, V, VI, or VII using an arene dioxygenase. These arene *cis*-dihydrodiols can then be subjected to chemical dehydration to obtain the corresponding trimethylphenols as shown in **Fig. 1**. Alternatively, 1,3,5-trimethylbenzene (Compound II) can be oxidized using an arene dioxygenase to obtain Compound VIII, which in turn can be dehydrated to obtain 2,4,6-trimethylphenol (Compound XII). In a presently preferred embodiment, an arene dioxygenase exhibits enhanced regiospecificity for the addition of the hydroxyl residues to the appropriate carbon atoms and/or enhanced enantiospecificity to obtain the desired chirality.

The invention also provides methods in which acyltransferases or glycosyltransferases are used to facilitate the production of a desired isomer of a dialkylphenol or a trialkylphenol from the *cis*-dihydrodiol intermediates that are formed upon arene dioxygenase-mediated biocatalysis. An example of this reaction is shown in **Fig. 3**, in which an acyltransferase or a glycosyltransferase is employed to acylate or glycosylate one of the cis-hydroxyl groups on a *cis*-dihydrodiol (Compound IV). Upon

chemical dehydration of the resulting compound (Compound XVI), the acylated or glycosylated group is preferentially lost, leaving a hydroxyl at the non-acylated position. By using an acyltransferase or glycosyltransferase that has been subjected to recombination and screening for those that are regiospecific for a particular hydroxyl group, using the methods of the invention, one can achieve specific production of the desired isomer of trimethylphenol (Compound IX).

Alternatively, one can employ chemical acylation in conjunction with an esterase to convert a *cis*-dihydrodiol di- or trialkylbenzene intermediate into a desired isomer of trialkyl or dialkylphenol. In this embodiment of the invention, the *cis*-dihydrodiol product of an arene dioxygenase reaction is subjected to chemical acylation with an anhydride, resulting in acylation of both hydroxyl groups (*e.g.*, Compound XVIII in **Fig. 3**). An esterase is then employed to release one of the acyl groups, thus producing the monohydroxyl derivative (*e.g.*, Compound XVI), which can then be converted to a desired dialkyl- or trialkylphenol by chemical dehydration (*e.g.*, Compound IX). In presently preferred embodiments, the esterase is a recombinant esterase that has been enhanced, using the methods of the invention, for improved properties such as regiospecificity and enantiospecificity, and the like.

Tocopherol derivatives that lack a methyl group, and thus use dimethylphenols as precursors, find use as antioxidants, among other uses. Such compounds can be synthesized using the methods, enzymes, and microorganisms of the invention. **Fig. 4** shows the various dimethylphenol compounds that one can produce by arene dioxygenase-catalyzed oxidation of xylenes, preferably in conjunction with whole cell biocatalysis, followed by chemical dehydration. For example, the invention provides methods in which *o*-xylene (Compound XXI) is oxidized by an arene dioxygenase to form one or more of the *cis*-dihydrodiols shown as Compounds XXV, XXVI, and XXVII. Chemical dehydration of Compounds XXV and XXVII can then be used to obtain 2,3-dimethylphenol (Compound XXXII) and 3,4-dimethylphenol (Compound XXXIV), respectively, while dehydration of Compound XXVI results in both of these compounds being produced. Accordingly, where a particular isomer is desired, in a presently preferred embodiment, the arene dioxygenase that is employed in the reaction is one that is optimized for the desired regiospecificity and/or enantiospecificity.

In additional embodiments, the invention provides methods in which an arene dioxygenase is used to catalyze the oxidation of *m*-xylene (Compound XXII) to one

or more of the arene *cis*-dihydrodiols Compound XXVIII, Compound XXIX, and Compound XXX. The arene *cis*-dihydrodiols can then, in turn, be subjected to chemical dehydration to obtain one or more dihydrophenols. For example, by dehydrating Compounds XXIX and XXX, respectively, 2,4-dimethylphenol (Compound XXXV) and 2,6-dimethylphenol (Compound XXXVI) are obtained. Dehydration of Compound XXVIII results in a mixture of the 2,4- and 2,6-dimethylphenols. Again, a particular isomer can be obtained by using an arene dioxygenase that has been optimized for the desired regiospecificity and/or enantiospecificity using the recombination and selection/screening methods of the invention.

p-Xylene (Compound XX) can also serve as a substrate for the arene dioxygenase-mediated oxidation. The resulting arene *cis*-dihydrodiols (Compounds XXIII and XXIV) can be chemically dehydrated to obtain 2,5-dimethylphenol. In these methods in particular, it is preferred that the arene dioxygenase is expressed by a cell that is of a species other than that from which the arene dioxygenase gene was obtained, or that the arene dioxygenase is expressed from a recombinant arene dioxygenase-encoding polynucleotide that has been optimized for improved properties using the recombination and selection/screening methods of the invention.

3. Preparation of hydroxylated aromatic carboxylic acids

Hydroxylated aromatic carboxylic acids have many diverse uses, including as antimicrobial additives, UV protectants (*e.g.* esters of p-hydroxybenzoic acid, parabens), pharmaceutical compositions (*e.g.*, esters of salicylic acid, coumarins and 3,4-dihydroxcoumarin).

Thus, in another preferred embodiment, the present invention provides a method for preparing hydroxylated aromatic carboxylic acids. The method involves contacting a substrate comprising an aryl carboxylic acid with a dioxygenase polypeptide of the invention. The polypeptide is preferably expressed by an organism of the invention.

a. Carboxylic acid substrates

The carboxylic acids used as substrates in the present invention can be obtained from commercial sources, or they can be prepared by methods known in the art. In a preferred embodiment, the carboxylic acids are prepared by contacting a substrate comprising an aryl alkyl group with an oxygenase polypeptide to produce the corresponding

aryl alkyl alcohol. The alcohol is subsequently acted upon by a dehydrogenase polypeptide to produce the desired carboxylic acid. Alternatively, the alcohol can be converted to COOH by chemical means.

For clarity of illustration, the discussion herein focuses on the oxidation of arylmethyl groups to carboxylic acids. This focus is intended to be illustrative and not limiting.

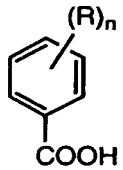
(i). Monooxygenation

The first step in the biotransformation processes for conversion of methylaryl compounds, such as toluene and isomeric xylenes involves the selective oxidation of at least one methyl group present in the aromatic substrate to the corresponding carboxylic acid (*e.g.*, benzoic, toluic acids). In an exemplary embodiment, the substrate is toluene, *p*- or, *m*- or *o*-xylene or 1,2,4-trimethylbenzene, or a mixture thereof, and preferably, only one of the methyl groups is oxidized.

Following the oxygenation step, the resulting alcohol is dehydrogenated, generally by the action of a dehydrogenase polypeptide to produce the desired carboxylic acid.

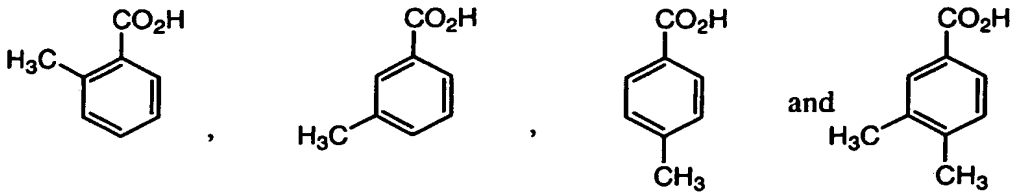
The invention provides for polypeptides that selectively oxidize only one alkyl group of an arene bearing two or more alkyl substituents. This embodiment is illustrated in **Fig. 6**, with the monooxidation of various xylenes. For example, p-xylene (2) is selectively converted to a monocarboxylic acid (22). Alternatively, the invention provides polypeptides that are capable of oxidizing more than one alkyl substituent of a species substituted with two or more alkyl groups. For example, certain polypeptides of the invention are capable of oxidizing both of the methyl substituents of a xylene, such as o-xylene (4) to the corresponding benzenedimethanol (4a).

In a preferred embodiment, the monoxygenation/dehydrogenation pathway produces a carboxylic acid having the structure:



wherein each of the n R groups is independently selected from H, alkyl and substituted alkyl groups; and n is an integer from 1 to 5, inclusive, more preferably R is methyl, and more preferably still, n is an integer from 1 to 3, inclusive.

In a still further preferred embodiment, the carboxylic acid group is selected from:



Many enzymes for effecting these reactions are well known in the art, and are suitable for use in the construction of useful polypeptides and host strains. To achieve the initial oxidation of the methyl groups, certain enzymes are presently preferred, including non-heme multicomponent monooxygenases of toluene and xylenes, and *p*-cymene, as well as certain arene dioxygenases which act on these substrate in a monooxygenase mode. The latter are exemplified by naphthalene dioxygenase, 2-nitrotoluene 2,3-dioxygenase and 2,4-dinitrotoluene 4,5-dioxygenase. These dioxygenases do not oxidize the aromatic ring of methylbenzenes, but are capable of oxidizing methyl groups of a variety of aromatic compounds in a monooxygenase mode (Selifonov, *et al., Appl. Environ. Microbiol,* **62**(2): 507-514 (1996); Lee *et al., Appl. Environ. Microbiol.* ;**62**(9):3101-3106 (1996); Parales, *et al, J.Bacteriol.,* **180**(5):1194-1199 (1998); Suen *et al, J.Bacteriol.*, **178**(16):4926-4934 (1996).

The following list provides examples of polynucleotides that encode a monooxygenase and are suitable for use in the methods of the invention. The loci are identified by GenBank ID and encode complete or partial protein components of the arene dioxygenases. Suitable loci include:

> [PSEXYLMA], [AF019635], [D63341], [E02361] xylene/toluene monooxygenase of *Pseudomonas putida* TOL plasmid (xyl M, xylA); [PPU24215] *p*-cymene monooxygenase of *P. putida*; [AF043544] nitrotoluene monooxygenase of *Pseudomonas* sp. TW3, NtnMA (ntnM, ntnA).

Additional monooxygenase polynucleotides useful in practicing the present invention are disclosed in USSN 09/373,928 entitled DNA SHUFFLING OF MONOXYGENASE GENES FOR PRODUCTION OF INDUSTRIAL CHEMICALS, to Joseph A. Affholter, Sergey A. Selifonov and S. Christopher Davis, Attorney Docket No.

018097-025810, filed on August 12, 1999, and incorporated herein by reference in its entirety.

In another preferred embodiment, the monooxygenase used is actually a dioxygenase that exhibits monooxygenase activity. As with the other polypeptide activities discussed herein, the ability of a dioxygenase to act as a monooxygenase is a property that can be optimized by shuffling the nucleic acids encoding these dioxygenases.

The following list provides examples of polynucleotides that encode dioxygenases acting as monooxygenases and which are suitable for use in the methods of the invention. The loci are identified by GenBank ID and encode complete or partial protein components of the arene dioxygenases. Suitable loci include:

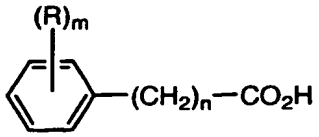
[AB004059], [AF010471], [AF036940], [AF053735], [AF053736], [AF079317], [AF004283], [AF004284], [PSENAPDOXA], [PSENAPDOXB], [PSENDOABC], [PSEORF1], [PSU49496] naphthalene-1,2-dioxygenase; [BSU62430] 2,4-dinitrotoluene dioxygenase; [PSU49504] 2-nitrotoluene dioxygenase.

A polypeptide that catalyzes monooxygenation can be a naturally occurring polypeptide, or it can have one or more properties that are improved relative to an analogous naturally occurring polypeptide. In a preferred embodiment, the polypeptides are expressed by one or more host organisms. Moreover, the polypeptide that catalyzes the monooxygenation can be co-expressed by the same host expressing a polypeptide used for further structural elaboration of the oxidation substrate or product (*e.g.*, a dioxygenase polypeptide that oxidizes the π-bond). Alternatively, the mono- and di-oxygenase polypeptides can be expressed in different hosts.

(ii). Oxidation of alkylarenes having alkyl groups with $\geq C_2$

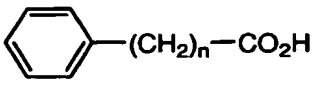
While much of the discussion above highlighting pathway and organism construction for oxidation of methylbenzenes is directly applicable to the set of processes dealing with alkyl benzenes bearing other alkyl group (**Fig. 7**), the essential distinguishing feature of the processes discussed in this section is in the provision of alternative enzymes for oxidation of these alkyl chains.

Thus, in a preferred embodiment, at least one alkyl group of the alkylarene has at least two carbon atoms. Preferred species produced in the monoxygenation step have the structure:



wherein each of the m R groups is selected from H, alkyl, substituted alkyl, aryl, substituted aryl, heteroaryl, substituted heteroaryl, heterocyclic and substituted heterocyclic; m is an integer from 0 to 5, inclusive; and n is an integer from 1 to 10, inclusive. Preferred aryl groups are those substituted on the aryl group with at least one methyl moiety.

In another preferred embodiment, the compound has the structure:



wherein n is an integer from 1 to 6, inclusive.

Certain exemplary embodiments are provided in **Fig.7** which illustrates the oxidation to a carboxylic acid of the terminal methyl groups of alkylbenzene compounds 5-9. The oxidation is accomplished by recruiting one or more genes encoding oxygenase activity. Generally, this is best accomplished by expressing a suitable cytochrome P450 type enzyme system. The enzymes of this class are ubiquitous in nature, and they can be found in a variety of organisms. For example, n-propylbenzene is known to undergo α-oxidation in strains of Pseudomonas desmolytica S449B1 and Pseudomonas convexa S107B1(Jigami *et al., Appl Environ Microbiol.* **38**(5):783-788 (1979)) which can utilize this hydrocarbon in either of two alternative oxidation pathways.

Similarly, well known in the art, alkane monooxygenases of bacterial origin, or cytochromes P450 for camphor oxidation, whether wild-type or mutant, can be recruited for the purpose of introducing the oxygen at the terminal methyl group of alkylarenes (Lee *et al.,Biochem. Biophys. Res. Commun.* **218**(1):17-21 (1996); van Beilen *et al., Mol. Microbiol.* **6**(21):3121-3136 (1992); Kok *et al., J. Biol. Chem.* **264**(10):5435-5441 (1989); Kok *et al., J. Biol. Chem.* **264**(10):5442-5451 (1989); Loida *et al., Protein Eng.* **6**(2):207-212 (1993).

(iii) Oxygenation of arenes with exocyclic π-bonds

In another preferred embodiment, the starting material for the carboxylic acid is an arene bearing an exocyclic π-bond. This class of compounds is exemplified by styrene. Other analogous species are set forth in **Fig. 11**.

The conversion of the exocyclic π-bond is best accomplished by recruiting a cluster of bacterial styrene oxidation genes well known in the art (Marconi *et al., Appl. Environ. Microbiol.* **62**(1):121-127 (1996); Beltrametti *et al., Appl. Environ. Microbiol.* **63**(6):2232-2239 (1997); O'Connor *et al., Appl. Environ. Microbiol.* **63**(11):4287-4291 (1997); Velasco *et al., J. Bacteriol.* **180**(5):1063-1071 (1998); Itoh, *et al., Biosci. Biotechnol. Biochem.* **60**(11):1826-1830 (1996). Alternatively, the styrene epoxidation step can be accomplished by using monooxygenases of methyl substituted aromatic compounds, such as toluene or xylenes (Wubbolts, *et al., Enzyme Micro.b Technol.* **16**(7):608-615 (1994).

(iv). Dehydrogenation

To produce the desired carboxylic acid, the alcohol from *(i-iii)*, above, is preferably treated with a dehydrogenase polypeptide. The dehydrogenase enzymes can be endogenous to a host that expresses one or more of the oxygenase polypeptides, or it can exhibit properties that are improved relative to an endogenously expressed dehydrogenase.

The polypeptide that catalyzes the dehydrogenation can be a naturally occurring polypeptide, or it can have one or more properties that are improved relative to an analogous naturally occurring polypeptide. In a preferred embodiment, the polypeptides are expressed by one or more host organisms. Moreover, the polypeptide that catalyzes the dehydrogenation can be co-expressed by the same host expressing one or more of the dioxygenase polypeptide. Alternatively, the dehydrogenase and oxygenase polypeptides can be expressed in different hosts.

In yet another preferred embodiment, the invention provides a method for altering or controlling the regiospecificity of the dehydrogenation reaction of a vicinal diol. This method "blocks" one of the vicinal diol hydroxyl groups by forming an ester, for example. The method involves contacting the vicinal diol with a polypeptide, preferably expressed by a host organism, having an activity selected from ligase, transferase and combinations thereof, thereby forming a α-hydroxycarboxylic acid adduct. As with the other polypeptides discussed above, this polypeptide can be expressed by the same host cell

that expresses other polypeptides of the reaction cascade. Moreover, this polypeptide can be a naturally occurring polypeptide, or it can be improved using the method of the invention.

The combined steps of oxygenation are illustrated in Fig. 5, with the preparation of compound 10 from compound 1, compound 22 from compound 2, compound 23 from 3 and compound 24 from compound 4.

b. Dioxygenation of aromatic π-bonds

In the synthesis of hydroxyaryl carboxylic acids using the methods of the invention, once the carboxylic acid moiety is in place, the molecule is submitted to a dioxygenation cycle. The dioxygenation of the aromatic ring is preferably accomplished by recruiting one or more arene dioxygenase genes, preferably of bacterial origin. Exemplary dioxygenase genes are disclosed herein. The method of the invention can be practiced using essentially any type of aromatic ring system. Exemplary aromatic systems include, benzenoid and fused benzenoid ring systems (*e.g.*, benzene, napthalene, pyrene, benzopyran, benzofuran, *etc.*) and heteroaryl systems (pyridine pyrrole, furan, *etc.*). In a preferred embodiment, the substrate includes a benzenoid hydrocarbon.

Similar to the embodiments discussed above, in this embodiment, the polypeptide that catalyzes the dioxygenation can be coexpressed with one or more polypeptides used in this synthetic pathway. For example, the monooxygenase, dehydrogenase and dioxygenase polypeptides can all be coexpressed in a single host. Other functional combinations of coexpression will be apparent to those of skill in the art.

In a preferred embodiment, benzoate-1,2-dioxygenase or toluate-1,2-dioxygenase are used to catalyze the formation of compound 14, p-cumate 2,3-dioxygenase to catalyze formation of compounds 13, 25, 26, and phthalate 4,5-dioxygenase or phthalate 3,4-dioxygenase to catalyze the formation of compound 12 (*see*, **Fig. 5**).

c. Exemplary embodiments

For the purpose of further illustration, methods of the invention for preparing coumarin derivatives and o-cinnamic acids are described below. These examples are intended to illustrate, not to limit, the scope of the present invention.

(i). Synthesis of 3,4-dihydrocoumarin

The present invention provides both a chemoenzymatic route to 3,4-dihydrocoumarin (43) from n-propylbenzene (steps as shown in **Fig. 8**), and means for the subsequent conversion of this compound to other lactone derivatives, such as coumarin (58) and 4-oxygenated derivatives of coumarin. While compound 43 can be converted to 58 by chemical methods known in the art (*e.g.* by reaction with sulfur, or by catalytic dehydrogenation over Pd or Pt catalyst), the purpose of this invention is also in the provision of alternative biocatalytic means for effecting such reaction.

In a preferred embodiment, one or more arene dioxygenase, such as naphthalene dioxygenase, toluene 2,3-dioxygenase or other structurally and functionally related dioxygenases, are shuffled, as described herein, to produce an improved polypeptide. In the method of the invention, the improved polypeptide is used to catalyze benzylic monooxygenation reactions with a variety of benzocycloalkanes, and benzylic desaturation reactions of compounds exemplified by 1,2-dihydronapthalene, indan, and ethylbenzene.

For the purpose of converting 3,4-dihydrocoumarin (43) to coumarin, these catalytic activities of the arene dioxygenase enzymes can be used to make either coumarin itself (desaturation), or 4-hydroxy-3,4-dihydrocoumarin (59). Regardless of enantiomeric composition, the latter compound can be chemically dehydrated (under acidic or heat conditions) to coumarin.

Alternatively, compound 59 can be microbially converted to acyl or glycosyl derivative 60 with the provision of corresponding transferase genes. When a dehydrogenase activity specific toward compound 59 is displayed by the host microbial strain, the variation of this route provides access to 4-hydroxycoumarin (70, R_8=H) (spontaneous tautomer of dehydrogenase product, 4-keto-3,4-dihydrocoumarin), or with provision of a suitable transferase gene to its 4-O-glycosyl, 4-O-Methyl or 4-O-acyl derivative 70. Availability of these routes is beneficial for making a range of other products originating in n-propylbenzene oxidative biotransformation pathway discussed above.

(ii). Conversion of naphthalene to coumarin and o-hydroxycinnamic acid

The invention also provides methods of making coumarin and o-hydroxy-cinnamic acids from naphthalene, using a whole-cell microbial biocatalyst, and methods for constructing microbial strains effecting such conversions in their entirety, or in part. In the

latter case, the last steps for making coumarin are accomplished by chemical methods. Exemplary reaction sequences used in these processes are shown in **Fig. 9**.

To construct an improved polypeptide, the invention preferably uses a subset of genes encoding catabolism of naphthalene by bacteria. In a preferred embodiment, at least four polypeptides (and the corresponding genes) are used to catalyze a series of reactions. These polypeptides include, naphthalene 1,2-dioxygenase (compound 10 to 61), NahA (a multicomponent enzyme); cis-1,2-dihydro-1,2-dihydroxynaphthalene dehydrogenase (compound 61 to 62), NahB; 1,2-dihydroxynaphthalene 1,1a dioxygenase (compound 62 to 63) and NahC.

Compound 63 is known to be labile, readily undergoing a series of tautomerization and intramolecular ring closure reactions. The *cis-trans* equilibrium between compounds 63 and 64 is preferably effected by 2-hydroxybenzalpyruvate isomerase (NahD) which can be used to impose a degree of control on the isomerization of the double bond. Preferably, NahE, 2-hydroxybenzalpyruvate hydratase/aldolase is preferably not used in this pathway, and preferred host strains are those essentially lacking activity of this enzyme.

In a preferred embodiment, the step of this process which allows for the preparation of either coumarin 58, or cis/trans o-hydroxycinnamic acids (66, 67) is the provision of an alpha-ketoacid decarboxylase enzyme with specific activity towards either compound 63 or 64 or both.

Several alpha-ketoacid decarboxylase enzymes are known, and in the preferred embodiment, the benzoylformate decarboxylase of *Pseudomonas putida*, or an enzyme structurally or functionally similar to it, is used (Gen Bank PSEMDLABC, benzoylformate decarboxylase (mdlC)).

Ring closure of compound 66 to 58 can be effected by enzymatic or chemical means (*e.g.* extraction under acidic conditions). Similarly, cis-trans isomerization of 67 to 66 can be effected by enzymatic or chemical means (*e.g.* by Pt or Pd- catalyzed hydrogenation/dehydrogenation under acidic conditions).

Biocatalytic variations of these processes, which can be used to produce coumarin from o-hydroxycinnamic acids, preferably involve the use of acyl-CoA ligase and transferase enzymes (pathway for conversion of 66 to 68 to 58), or conversion of 67 to glycoside 69 with subsequent isomerization later in an enzymatic process akin to that of coumarin biosynthesis in plants.

4. Preparation of α-hydroxycarboxylic acids

α-hydroxycarboxylic acids (AHAs) are an important group of industrial chemicals. One of the simplest representatives of this class of compounds is lactic acid which find many uses, including synthesis of polyester polymers (polylactic acid). Other representative AHAs, such as mandelic acid can also be used as a constituent of polymers or co-polymers with lactic acid. Enantiomerically pure AHAs are also used as resolving reagents for separating racemates of chiral molecules.

AHAs are typically generated chemically by hydrolysis of a cyanohydrin, generally prepared from an aldehyde. Aldehydes, however, are relatively expensive starting materials, and the cyanohydrin pathway does not readily provide for direct preparation of AHAs in high enantiomeric excess. One pathway for the synthesis of AHAs in high enantiomeric excess is through the use of one or more enzymatic reactions starting from an inexpensive and readily available starting material such as an alkene.

Arene dioxygenases (ADOs) are known to oxidize alkenes to the corresponding vicinal diols. ADOs, such as toluene 2,3-dioxygenase, isopropylbenzene 2,3-dioxygenase, benzene-1,2-dioxygenase, biphenyl-2,3-dioxygenase naphthalene-1,2-dioxygenase, and many homologous and/or functionally similar enzymes can be used to manufacture AHAs in a highly regioselective fashion. An example of this dioxygenation reaction is provided in **Fig. 12**, with the conversion of alkene (I) to the vicinal diol (II).

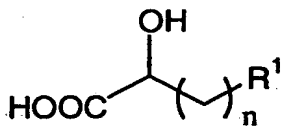
The present invention provides general methods for the biocatalytic manufacture of AHAs and their esters, and also for the construction and optimization of the biocatalytic properties of enzymes and host strains that effect oxidative cis-dihydroxylation or epoxidation reactions of a variety of alkenes. Moreover, the invention provides for the subsequent enzymatic conversion of the dioxygenation products to AHAs and ester derivatives of AHAs.

Thus, in another preferred embodiment, the invention provides a method for converting an olefin into an α-hydroxyacid. The method involves: (a) contacting the olefin with an improved dioxygenase polypeptide to form a vicinal diol; and (b) contacting the vicinal diol with a dehydrogenase polypeptide to form the α-hydroxyacid.

The polypeptide that catalyzes the dehydrogenation can be a naturally occurring polypeptide, or it can have one or more properties that are improved relative to an analogous naturally occurring polypeptide. In a preferred embodiment, the polypeptides are

expressed by one or more host organisms. Moreover, the polypeptide that catalyzes the dehydrogenation can be co-expressed by the same host expressing the dioxygenase polypeptide that oxidizes the π-bond. Alternatively, the dehydrogenase and dioxygenase polypeptides can be expressed in different hosts. An example of the dehydrogenation is provided in **Fig. 12**, with the conversion of diol (II) to aldehyde (III) and its conversion to α-hydroxycarboxylic acid (IV). As shown in **Figs. 14A-14R**, this same method can be carried out on a substrate comprising an aromatic moiety. The exemplary two-step dehydrogenation can equally well be carried out using a one step process.

Although the method of the invention can be used to produce AHAs having substantially any structure, in another preferred embodiment, the α-hydroxycarboxylic acid has the structure:



wherein R^1 is selected from aryl, substituted aryl, heteroaryl, substituted heteroaryl, heterocyclic, substituted heterocyclic, —NR^2R^3, —OR^2, —CN, $C(R^4)NR^2R^3$ and $C(R^4)OR^2$ groups; R^2 and R^3 are members independently selected from H, alkyl, substituted alkyl, aryl, substituted aryl, heteroaryl, substituted heteroaryl, heterocyclic and substituted heterocyclic groups; R^4 is selected from =O and =S, and n is an integer between 0 and 10, inclusive.

In a further preferred embodiment, R^1 is selected from phenyl, substituted phenyl, pyridyl, substituted pyridyl, —NR^2R^3, —OR^2, —CN, $C(R^4)NR^2R^3$ and $C(R^4)OR^2$ groups, R^2 and R^3 are members independently selected from H, alkyl, substituted alkyl, aryl, substituted aryl, heteroaryl, substituted heteroaryl, heterocyclic and substituted heterocyclic groups; and R^4 is selected from =O and =S.

a. <u>α-Hydroxycarboxylic acid adducts</u>

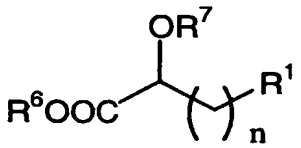
AHAs are bifunctional molecules with two chemically and enzymatically distinguishable functional groups, carboxyl and hydroxyl. In the biocatalytic modifications of AHAs described in this invention, either of these groups can be derivatized by bond formation. While these reactions do not change the oxidation state of the AHA molecule, recruitment of the enzymes effecting modification of AHAs provides the opportunity to generate biotransformation endproducts with substantially different physical and chemical properties than that of a free AHA. Generally desirable properties include an increase of

hydrophobicity, a decrease of aqueous solubility and, for an ester formed through a carboxylic group of an AHA, a decrease in acidity of the process end-products.

In a preferred embodiment, the adduct-forming polypeptide produces an α-hydroxycarboxylic acid adduct selected from esters and ethers. The method involves contacting an α-hydroxycarboxylic acid with a polypeptide having an activity selected from ligase, transferase and combinations thereof, thereby forming a α-hydroxyacid adduct. The adduct forming polypeptides useful in this embodiment can be naturally occurring polypeptides or, alternatively, they can be polypeptides improved using the methods of the invention, as discussed generally, above.

Exemplary adduct forming reactions are provided in **Fig. 13**. This Figure shows the use of a methyltransferase to convert carboxylic acid (X) to the corresponding methyl ester (XI), acyltransferase I to convert compound X to ester XIII, and acyl-CoA ligase to convert X to intermediate XIV. This intermediate can then be transformed into a simple alkyl ester (XIX) or to structures having greater complexity of structure in the alcohol-derived component (*e.g.*, XV). Species such as XV can be further elaborated using other polypeptides including, for example, acyltransferase III to produce compound XVII, thioesterase II to produce compound XVIII and thioesterase I to produce compound XVI.

In a further preferred embodiment, the α-hydroxycarboxylic acid adduct has the structure:



wherein, R^1 is selected from aryl, substituted aryl, heteroaryl, substituted heteroaryl, heterocyclic, substituted heterocyclic, —$NR^2R^3(R^4)_m$, —OR^2, —CN, $C(R^5)NR^2R^3$ and $C(R^5)OR^2$ groups, R^2, R^3 and R^4 are members independently selected from the group consisting of H, alkyl, substituted alkyl, aryl, substituted aryl, heteroaryl, substituted heteroaryl, heterocyclic and substituted heterocyclic groups; R^5 is selected from =O and =S; R^6 is selected from H, alkyl and substituted alkyl groups; R^7 is $C(O)R^8$, wherein R^8 is selected from H alkyl and substituted alkyl groups and R^7 and R^8 are not both H; m is 0 or 1, such that when m is 1, an ammonium salt is provided; and n is an integer between 0 and 10, inclusive.

In yet another preferred embodiment, R^1 is selected from phenyl, substituted phenyl, pyridyl, substituted pyridyl —NR^2R^3, —OR^2, —CN, $C(R^5)NR^2R^3$ and $C(R^5)OR^2$ groups; R^2 and R^3 are members independently selected from the group consisting of H, C_1-C_6 alkyl and allyl; and R^5 is =O.

In yet another preferred embodiment of this invention, the described reactions and pathways are utilized for biocatalytic whole-cell conversion of styrene to mandelic acid and its ester derivatives. The pathway for styrene conversion, all of its intermediates and reactions are shown in **Figs. 14A- 14R**.

The esterified adducts provide an increase in the overall efficiency of the biotransformation process as they simplify end-product recovery. The esters are easily isolated by organic solvent extraction and partitioning. Moreover, the adducts obviate the need for pH adjustment in the aqueous fermentation media to prevent the accumulation of the high levels of acidic biotransformation products.

There are several biochemically distinct means by which AHAs can be biocatalytically esterfied in a substantially aqueous environment. In one preferred embodiment of this invention, expression of genes encoding an S-adenosylmethionine (SAM)-dependent O-methyltransferase is used to effect conversion of AHAs to their methyl esters (*e.g.*, **Fig. 13**, conversion of compound X to compound XI). SAM-dependent methyltransferases of differing substrate specificity are common in nature, and suitable enzymes and corresponding genes can be found and used directly for the purpose of this invention. Alternatively, these species can be further evolved and optimized for specific activity with the AHAs using one or more nucleic acid shuffling methods described herein. The invention also provides means for HTP screening for the presence, and quantitative determination, of the AHA-specific O-methyltransferase catalytic activities in microorganisms, cells, tissues or extracts of tissues of higher eukaryotic organisms. These methods can be used either to identify sources of corresponding genes or to evolve the desired specificity of known methyltransferases towards the AHAs by nucleic acid shuffling as described herein.

In another embodiment acyltransferase enzymes which specifically esterify the sec-hydroxyl of AHAs by means of active carboxyl transfer from either acyl-coenzyme A or acylated acyl carrier protein (ACP) are incorporated into the reaction pathway. This pathway is depicted in **Fig 13**, as shown by the coupling of compounds X and XII to yield compound XIII. A preferred embodiment of this pathway, involves recruiting and

expressing gene(s) encoding acyl-CoA-dependent acyltransferases, including those which utilize as substrates acetyl-CoA and CoA derivatives of fatty acids, as well as lactoyl-CoA, CoA-thioesters with other AHAs, and CoA derivatives of aromatic, arylalkanoic, branched chain alkanoic carboxylic acids, and alpha-aminoacids. Where carboxylic acids (either in the form of a free acid, salt or ester), intended for esterification of AHAs, are supplied exogenously, or are co-produced by another co-functioning biotransformation or fermentative pathway in the same host organism, or a different host organism, the invention provides a means for facilitating ester formation by recruiting and co-expressing those acyl-CoA ligases or ACPs which effect *in-vivo* activation of these acids forming suitable substrates for the acyl transferase enzymes that act on the AHAs.

The invention also provides for another type of biochemical transformation of AHAs to AHA carboxylic esters wherein free AHAs are first converted to their active ester form by means of the enzymatic formation of a derivative with CoA or ACP (**Fig. 13**, compound XIV). Several alternative acyltransferase enzymes (and genes encoding them) can be recruited for effecting subsequent transformations of compound XIV to esters of different compositions. These preferably include AHA-CoA transferases acting (*a*) on alcohols (XX) to produce esters, or (*b*) on molecule of compound XIV or compound XV to produce acyclic homo- and hetero- oligomers (n=2-5) of AHAs. By recruiting an additional thioesterase enzymes, the activated forms of these oligomeric esters can be converted to free carboxylic oligomers (*e.g.*, XVIII) or to the cyclic substituted glycolides (XVI).

In another preferred embodiment, the formation of an α-hydroxycarboxylic acid ester is catalyzed by an acyl CoA-ligase that is evolved by nucleic acid shuffling. In a preferred embodiment, shuffling of nucleic acids encoding acyl-CoA ligase activities results in an increase in the synthesis of esters. In another preferred embodiment, the esters are selected from structures XIII-XVIII (**Fig. 13**). The synthesis of these and other esters will generally rely on the provision of a corresponding α-hydroxycarboxylic acid precursor. In a preferred embodiment, the α-hydroxycarboxylic acid precursor is present in an amount sufficient to establish intracellular pools of CoA-activated carboxylic derivatives of α-hydroxycarboxylic acids.

In still another preferred embodiment, the transferase polypeptide is selected from glycosyltransferase and methyltransferase, more preferably methyltransferase and more preferably still a S-adenosylmethionine dependent O-methyltransferase.

5. Enzymes effecting chiral switch at the level of AHAs.

Another object of this invention is the effective control of the enantiomeric composition of the compounds prepared by the methods of the invention. For clarity of illustration, the discussion below focuses on AHA esters made by the biotransformation process from alkenes. This focus is intended to be illustrative and not limiting of the scope of this embodiment of the invention.

Means of enantiomeric control, when integrated as part of the multistep biocatalytic pathway, constitutes an important advantage as it allows selective production of either enantiomer of the AHA. The enantiomerically pure AHAs can be used as resolving reagents, chiral synthons, or monomers for polyesters or co-polyesters with lactic acid.

In a preferred embodiment, the AHA is mandelic acid, or an analogue thereof, and the chiral switch is effected by recruiting mandelate a racemase gene.

Mandelate reacemase catalyzes the interconversion of the R and S enantiomers of mandelic acid and its derivatives. An exemplary mandelate racemase is that of *Pseudomonas putida* (the sequence of the gene can be found in the GenBank database under the locus [PSEMDLABC]). Preferred mandelate racemases are those of the *P. putida* strain ATCC 12633, however, mandelate racemases from any other organism can be used.

Although, in a preferred embodiment, the chiral switch is made at the level of the AHA, this switch can be made with any of the precursors or adducts of the AHA as well. Thus, in yet another preferred embodiment, the AHA is modified by at least one of the ester-forming enzymes discussed herein. Preferred ester forming enzymes are those which specifically, or preferentially, act on one enantiomer of the AHA, thus allowing enantiospecific resolution of the racemate *in-vivo*. The activity of the above racemases provides an enantiomeric equilibrium at the expense of the non-esterified enantiomer. The combined action of the racemase and the AHA esterifying enzymes provides a chiral switch which allows preparation of one desired enantiomer, whether R or S, from AHAs of any enantiomeric composition.

E. Antioxidant and Impurity Modification and Detoxification

In another embodiment, the invention provides methods of degrading or modifying organic materials which lead to their detoxification. Exemplary compounds include stabilizing agents, antioxidizing agents, environmental pollutants and the like. This method is applicable to substantially any compound that can be detoxified by, for example, oxidation, either with or without additional structural elaboration. For clarity of illustration,

the discussion below focuses on the detoxification of agents commonly found in organic solvents and in π-bonded compounds of use in the present invention.

Many commercially available compounds (*e.g.*, alkylbenzenes, alkenes, *etc.*) are stabilized with small amounts of antioxidants such as 4-*tert*-butylcatechol or alkylphenols (*e.g.* BHT) to prevent polymerization during storage and transportation. While the amount of these compounds is usually relatively small (10-15 ppm), they can inhibit biocatalyst performance as they accumulate in aqueous fermentation medium during prolonged incubations required to obtain satisfactory endproduct concentrations.

Several types of enzymes for modifying the phenolic stabilizing compounds can be used to alleviate any negative effects of these compounds on the whole cell biocatalyst performance. Their genes can be introduced in the same host organism used to produce endproducts or intermediates of relevance to this invention. Alternatively, they can be incorporated into a separate host organism. This obviates any need for additional steps in the process to remove these stabilizers. Optimization of one or several of these enzymes for the efficient removal of these stabilizing compounds is a target for nucleic acid shuffling.

Exemplary enzymes for modifying phenolic and diphenolic stabilizers include, but not limited to, acyltransferase, methyltransferase, glycosyltransferase, lactase and peroxidase. In addition to these enzymes, catecholic stabilizers also can be modified to innocuous products by catechol dioxygenases effecting *meta*- or *ortho*-ring cleavage. Many of these enzymes show a significant breadth of activity towards compounds related to phenolic stabilizers. Thus, nucleic acid shuffling can be applied to optimize enzyme parameters such as:

a) increased turnover with particular phenolic stabilizer,

b) increased functional expression, by obviating the requirements for certain post-transitional modifications of those enzymes which require such modifications (*e.g.* glycosylation of peroxidases and lactases); and

c) alleviation of inhibition of these enzymes by high concentration of co-occurring feedstock compounds and intermediates and endproducts of the biocatalytic process.

F. Analytical Methodology

A number of analytical techniques are useful in practicing the present invention. These analytical techniques are used to measure the extent of conversion of a particular substrate to product. These techniques are also used to analyze the regioselectivity and/or the enantiomeric selectivity of a particular reaction catalyzed by a polypeptide of the invention. Moreover, these techniques are employed to assess the effect of nucleic acid shuffling experiments on the efficiency and selectivity of the polypeptides produced following the shuffling. The discussion below focuses on those aspects and embodiments of the invention in which an olefin precursor is oxidized by a dioxygenase. The analytical techniques discussed in the this context are generally of broad applicability to other aspects and embodiments of the invention. This is particularly true of the spectroscopic and chromatographic methods discussed below. Thus, in the interest of brevity, the following discussion focuses on analyzing the products of the oxidation of an olefin, but the utility of the methods discussed is not limited to this embodiment.

The generation and screening of high quality shuffled libraries provides for nucleic acid shuffling (or "directed evolution"). The availability of appropriate high-throughput analytical chemistry to screen the libraries permits integrated high-throughput shuffling and screening of the libraries to achieve a desired dioxygenase activity. Moreover, this discussion is also generally applicable to those dioxygenases that act as monooxygenases.

1. Selecting for Dioxygenase activity

Dioxygenase activity can be monitored by HPLC, chiral HPLC, gas chromatography, NMR spectrometry, and mass spectrometry, as well as a variety of other analytical methods available to one of skill. Incorporation of ^{18}O from radiolabeled molecular oxygen can be monitored directly by mass shift by MS methods and by an appropriate radioisotope detector with HPLC and GC devices. For example, oxidation of 1-hexadecene to 1,2-hexadecanediol can be monitored by ^{18}O incorporation either in intact whole cells or lysate. This has been used, for example by Bruyn et al. with *Candida lipolytica.*

In addition, epoxide formation can be indirectly measured by various reactive colorimetric reactions. When H_2O_2 is used as the oxidant, disappearance of peroxide over time can be monitored directly either potentiometrically or colorimetrically using a number of commercially available peroxide reactive dyes.

In a high-throughput modality, a preferred method is high-throughput MS, or MS operating in a coordination ion spray and/or electrospray-based mode. In addition, selection protocols in which the organism uses a given alkene or aromatic system as a sole carbon source can be used. In some systems this will be most readily accomplished by using the dioxygenase to generate a metabolizable diol.

2. Automation for Strain Improvement

One key to strain improvement is having an assay that can be dependably used to identify a few mutants out of thousands that have potentially subtle increases in product yield. The limiting factor in many assay formats is the uniformity of library cell (or viral) growth. This variation is the source of baseline variability in subsequent assays. Inoculum size and culture environment (temperature/humidity) are sources of cell growth variation. Automation of all aspects of establishing initial cultures and state-of-the-art temperature and humidity controlled incubators are useful in reducing variability.

In one aspect, library members, *e.g.*, cells, viral plaques, spores or the like, are separated on solid media to produce individual colonies (or plaques). Using an automated colony picker (*e.g.*, the Q-bot, Genetix, U.K.), colonies are identified, picked, and 10,000 different mutants inoculated into 96 well microtitre dishes containing two 3 mm glass balls/well. The Q-bot does not pick an entire colony but rather inserts a pin through the center of the colony and exits with a small sampling of cells, (or mycelia) and spores (or viruses in plaque applications). The time the pin is in the colony, the number of dips to inoculate the culture medium, and the time the pin is in that medium each effect inoculum size, and each can be controlled and optimized. The uniform process of the Q-bot decreases human handling error and increases the rate of establishing cultures (roughly 10,000/4 hours). These cultures are then shaken in a temperature and humidity controlled incubator. Glass or, preferably, stainless steel balls in the microtiter plates act to promote uniform aeration of cells and the dispersal of mycelial fragments similar to the blades of a fermenter.

a. Prescreen

The ability to detect a subtle increase in the performance of a shuffled library member over that of a parent strain relies on the sensitivity of the assay. The chance of finding the organisms having an improvement is increased by the number of individual mutants that can be screened by the assay. To increase the chances of identifying a pool of sufficient size, a prescreen that increases the number of mutants processed by 10-fold can be

used. The goal of the primary screen will be to quickly identify mutants having equal or better product titres than the parent strain(s) and to move only these mutants forward to liquid cell culture for subsequent analysis.

For the purpose of preparing a shuffled ADO library screening and sorting out non-functional variants, several general activity detection methods for ADOs can be used, including cases for direct screening and colony picking on agar medium plates.

Certain presently preferred examples of general methods are:

(a) the formation of indigo from indole, and similarly, from substituted indoles and indole-carboxylic acids (to produce indigo or substituted indigo). The development of blue or blue-grey hued growing colonies signifies the expression of catalytically functional ADOs. Most of the ADOs exhibit some activity with either indole (*e.g.* toluene dioxygenase, biphenyl dioxygenase, naphthalene dioxygenase and homologous enzymes), or with indole carboxylates (e.g. toluate-1,2-dioxygenase, p-cymate 2,3-dioxygenase);

(b) the detection of catechol formation, which can be enhanced in the presence of an aromatic amine (*e.g.* p-toluidine) and iron salts. Many catechols oxidize readily under oxygen, or in the presence of other oxidants, forming various colored products in the media surrounding the colonies expressing catalytically active ADOs.

This assay method is applicable for cases where a cis-diol is unstable (angular dihydroxylation products at aromatic π-bonds substituted with heteroatoms such as N, O, S, and halogens) and rearomatized spontaneously with concomitant elimination of a leaving substituent. The leaving substituent (*e.g.* halide, nitrite or ammonia) can also be detected by a variety of qualitative assays known in the art, or by detecting changes in pH of surrounding medium (pH indicator).

The assay can also be used in cases of stable dihydrodiol formation, where a suitable gene of arene cis-diol dehydrogenase is co-expressed (in many cases such genes are indeed available and well known in the art). In this case, the accumulated catechol can be detected by the presence of colored oxidation products, whether enhanced by p-toluidine/Fe or not; and

(c) color formation due to enzyme activities encoded by accessory genes for subsequent metabolism of arene cis-dihydrodiols. In this case, a yellow (or orange) color is developed by colonies in the medium which signifies the expression of catalytically active ADO.

A combination of the methods can be used. The advantage of the indole/indigo assay is that the color does not diffuse into areas of medium surrounding the colonies. The advantage of other assays is that they often can provide information about both the general activity and also about the regioselectivity of the reaction catalyzed by ADO.

Positive clones selected by either of the above-described exemplary methods can be examined by subsequent tier assay.

3. Screening for improved dioxygenase activity.

In each of the aspects and embodiments discussed herein, the concept of screening the library of recombinant polypeptides to enable the selection of improved members of the library is set forth. Although it will be apparent to those of skill in the art that many screening methodologies can be used in conjunction with the present invention, the invention provides a screening process comprising:

(a) introducing the library of recombinant polynucleotides into a population of test microorganisms such that the recombinant polynucleotides are expressed;

(b) placing the organisms in a medium comprising at least one substrate; and

(c) and identifying those organisms exhibiting an improved property compared to microorganisms without the recombinant polynucleotide.

a. Oxidation of olefins

Depending on the specific outcome desired from a particular course of shuffling of nucleic acids encoding oxygenases for biocatalytic oxidation of olefins, the invention provides several methods for detecting and measuring catalytic properties encoded by the recombinant polynucleotides. These are exemplified by the following methods.

Optimizing individual reactions and whole pathways for producing oxidized compounds, their derivatives, analogues and precursor compounds described in this invention can be monitored by virtually any analytic technique known in the art. In preferred embodiments, the production of the desired compound is monitored using one or more techniques selected from thin layer chromatography (TLC), high performance liquid chromatography (HPLC), chiral HPLC, mass-spectrometry, mass spectrometry coupled

with a chromatographic separation modality, NMR spectroscopy, radioactivity detection from a radioactively labeled compounds (*e.g.*, olefins, diols, carboxylic acids, aldehydes, AHAs, *etc.*), scintillation proximity assays, and by UV-spectroscopy. In a high throughput modality, the preferred methods are selected from one or any combination of these methods.

(i). Methods for measuring glycol formation from π-bonded species

The methods of the invention are used to improve polypeptides that catalyze the initial oxidation of π-bonded species. Methods using dioxygenase-based pathways are encompassed herein. The oxidation product from the conversion of a substrate comprising a π-bond (*e.g.*, arenes, alkylarenes, alkenes, *etc.*) can be detected by numerous methods well known to those of skill in the art. Certain preferred methods are set forth herein.

In a preferred embodiment, the vicinal diol derived from oxidation of an olefin is quantitated using a radioactively labeled substrate. Although any radioactive isotope commonly used in the art can be incorporated into a substrate, preferred isotopic labels include, for example, ^{14}C and/or ^{3}H. Differences in the volatility of the olefin substrate and the corresponding diol can be exploited to quantitate the radioactively labeled product. This method can easily be applied to aqueous samples of culture fluids obtained by incubating individual clones of cells expressing libraries of a recombinant polynucleotide obtained using the methods of the invention.

In an exemplary embodiment, cells expressing libraries of recombinant polynucleotides encoding a dioxygenase can be grown in a multiwell dish with a radioactive substrate administered directly to the aqueous medium. After incubation of the cells with the radioactive olefin substrate, any residual uncoverted substrate is removed by evaporation, with or without application of vacuum. After removing the unconverted substrate, the culture fluid (or aliquots thereof) is mixed with a suitable scintillation cocktail, and the radioactivity in the samples is quantitatively measured. In a preferred embodiment, selection of the most active clones is based on the amount of radioactivity incorporated into the compounds produced by the organisms expressing the clone.

Alternatively, radioactively labeled substrate can be administered as a vapor phase to colonies growing on a surface of a membrane filter overlaying agar-solidified medium. After incubation, the membrane is removed from the agar surface, and any residual hydrocarbon is evaporated from the membrane. The membrane is autoradiographed, or a scintillation dye is sprayed over the membrane for radioactivity

detection. A modification of this assay that is particularly suitable for ^{14}C label detection in and/or around colonies capable of oxidizing π-bonds to the corresponding glycols involves using a porous membrane that has scintillation dye incorporated in the membrane composition by covalent or adsorption means. This assay is termed "scintillation proximity assay on membrane" or "SPA."

In another embodiment of this invention, a variation of SPA is used to selectively quantify the glycol derived from the substrate. This variation involves adding beads for scintillation proximity assay to the samples of culture fluids or extracts obtained by incubation of cells with radiolabeled substrate as described above. Alternatively, the sample can be applied to a membrane. The beads or membrane are functionalized with groups that interact with a glycol.

In a preferred embodiment of this assay, the beads or membranes contain a suitable scintillating dye and their surfaces are modified by chemical groups that interact readily with diols. Such materials can be prepared by known chemical methods from commercially available SPA materials and they can be used to trap free diols directly in the aqueous medium or culture broths obtained by incubation of the microbial cells with the radiolabeled substrates.

In another preferred embodiment, the surface of the beads used in this assay is functionalized with a sufficient amount of a compound that interacts with a glycol, such as compounds containing aryl or alkylboronate (boronic acid). Such beads can be obtained by chemical modification of commercially available SPA beads by reactions known to one skilled in the art. In a preferred embodiment, the reactions used to modify the beads are analogous to those used for the preparation of arylboronate-modified resins for solid-phase extraction or chromatography. After incubation, the beads are washed with a sufficient amount of water or other suitable solvent and subjected to quantitative determination of radioactivity.

One can also determine amounts of glycol produced by oxidation of an π-bond by taking advantage of the reactive nature of the substrate. Samples of culture fluids, or extracts in an appropriate solvent, can be treated with known excess amounts of dilute solutions of, for example, a halogen (Cl_2, Br_2, I_2), permanganate salts. The residual excess amount of those reagents, left after reaction with any substrate present, can be measured by chemical methods known in the art for determination of these compounds (*see*, for example,

VOGEL'S PRACTICAL ORGANIC CHEMISTRY 5th Ed., Furniss *et al.*, Eds., Longman Scientific and Technical, Essex, 1989).

Mass spectrometry can also be used to determine the amount of a vicinal glycol formed due to species encoded by the libraries of shuffled oxygenase genes. Mass spectrometric methods allow ion peaks to be detected. The ion peaks derived from the vicinal glycol can be readily distinguished from peaks derived from olefin substrates. In a preferred embodiment, coordination ion spray or electrospray mass spectrometry is utilized.

In another preferred embodiment, a compound that interacts with a component of the mixture, preferably the glycol, is utilized to enhance the sensitivity and selectivity of the method. In a presently preferred embodiment, the sample analyzed contains excess arylboronic or alkylboronic acid. Preferred boronic acids are those containing at least one nitrogen atom and include, but are not limited to, dansylaminophenylboronic acid, aminophenylboronic acid, pyridylboronic acid.

The ions detected in the mass spectrum derive from cyclic boronate ester derivatives of the glycols with a boronic acid. The samples are preferably analyzed in non-acidic and non-basic organic solvent or aqueous phase, substantially free of alcohols and other glycols. Other appropriate analytical conditions will be apparent to those of skill in the art.

Another preferred method for quantitating the glycols uses periodic acid or its salts, preferably the sodium salts, to cleave the vicinal glycols to the corresponding aldehydes. In a preferred embodiment, vicinal diols other than the analyte (*e.g.*, carbohydrates) are excluded from the aqueous or organic solvent samples. This is easily attained by using non-carbohydrate carbon sources to grow the microbial cells, and/or by removal of the cells from the media by centrifugation or filtration prior to contacting of the sample with periodate reagent. The periodate reagent can be used in solution, or preferably, immobilized on a solid phase (*e.g.* anion exchange resin). After reacting the glycol with an excess of periodate ion, the amount of free aldehyde groups can be measured by a variety of assays know in the art. In a preferred method, the aldehydes are quantitated by a method based on the formation of a colored hydrazone derivative. Alternatively, when using radioactively labeled olefins for biotransformation, the free aldehydes obtained by this method can be trapped by aldehyde reactive groups (*e.g.*, free amines) on the surface of an appropriately modified SPA beads or membranes.

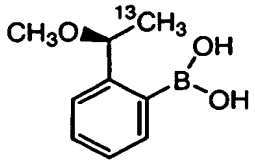
(ii). Methods for detecting alternative regioselectivity of oxidation of species with multiple π-bonds

In one embodiment, the substrate includes more than one π-bond (*e.g.*, styrene, butadiene, *etc.*). In a preferred embodiment, one of the π-bonds undergoes reaction more readily than the other. In this embodiment, it is generally preferred to determine which of the π-bonds underwent reaction. The preferred method for making this determination is ^{1}H or ^{13}C NMR, although other methods can be used. Other methods include, for example, chromatography (*e.g.*, TLC, GC, HPLC, *etc.*), UV/vis spectroscopy and IR spectroscopy. In an embodiment wherein the reaction is operating in a high throughput mode, the method of choice is flow-through ^{1}H or ^{13}C NMR spectroscopy.

When ^{13}C NMR is used, the substrates are preferably labeled with ^{13}C. π-bonded species can be synthesized by methods know in the art from a ^{13}C enriched material to incorporate one, or any combination of several, labeled carbon atom(s) into the structure of these compounds (by synthetic methods known in the art as exemplified by Selifonov *et al, in Appl Environ Microbiol.* **64**(4):1447-53 (1998)). The enrichment levels for the labeled positions are preferably at least 5% of ^{13}C, more preferably 50% and more preferably still 95% for any given labeled position. Incorporation of a ^{13}C label provides a number of advantages, such as increasing the NMR signal and decreasing time required for spectral acquisition. Moreover, labeled compounds allow for a quantitative or semi-quantitative interpretation of the composition of a mixture of isomeric oxidation products. Preferably, incubations with ^{13}C labeled olefins are conducted in multi-well plates, and aliquots of culture fluids or their extracts are sampled with an autosampler communicating with the NMR probe. In another preferred embodiment, the reaction components are not chromatographed or otherwise purified prior to obtaining a NMR spectrum.

Determining the absolute configuration and the enantiomeric composition of the glycols formed from π-bonded species, preferably employs a variation of the method described above for determining regioselectivity of dihydroxylation of the olefinic substrates by a dioxygenase using ^{1}H or ^{13}C NMR. In a preferred embodiment, the substrates are labeled with ^{13}C and ^{13}C NMR, is employed. This method preferably involves the use of a chiral and essentially enantiomerically pure derivatizing reagent such as a substituted arylboronic acid which forms a cyclic boronate derivatives with vicinal glycols, as know in the art (Burgess and Porte, *Angew. Chem Intl Ed. Engl.* **33**(41):1182-4

(1994)). In a preferred embodiment, both the substrates and one or more carbon atoms of the boronic acid is labeled with ^{13}C. Although a broad range of boronic acids are of use in the present invention, a currently preferred boronic acid is shown below:



The absolute configuration of any chiral center of the compounds produced by the methods of the invention can be either R or S. In presently preferred embodiments, the enantiomeric excess of the product is preferably 98% or more. NMR signals of different enantiomers of the reaction products can be distinguished in diastereomeric products using substantially enantiomerically pure boronate compounds as discussed above. Moreover, the relative intensity of the NMR signals arising from corresponding atoms of the diastereomeric products can be used for estimating the enantiomeric composition of the product(s) present in the sample.

(iii). Methods for detecting alternative regioselectivity of oxidation of alkylarenes

Useful methods for determining the regioselectivity of the oxidation of alkylarene compounds are substantially similar to those described in section (ii), *supra.*

4. AHA formation from glycols

Among methods for specifically measuring the free AHAs produced in the biocatalytic process, those which are particularly preferred are methods using a variation of the scintillation proximity assay described above. These methods preferably use an excess of beads or membranes bearing one or more positively charged functional groups (*e.g* quaternary or tertiary or primary amines). In preferred embodiments, these beads or membranes act as an anion exchange medium and they selectively trap free AHAs, thereby removing them from aqueous culture broths. In another preferred embodiment, this method employs a radioactively labeled starting material, or subsequent intermediate, (*e.g.*, glycol, epoxide, *etc.*). The radioactively labeled compound interacts with the beads or membrane. Prior to measuring the radioactivity associated with the beads or the membrane, non-specifically adsorbed label is preferably removed by evaporating excess radioactive

compound and/or washing with an aqueous solution which does not cause elution of the AHAs from the anion-exchange beads or membrane.

Preferred methods for determining the chirality and absolute configuration of AHAs formed in the described biotransformation process are substantially similar to those methods employed in making these determinations with respect to the glycols, as discussed above.

5. Methods for determination of HCAs

In HTP mode, a preferred analytical method is flow-through ^{1}H or ^{13}C NMR spectroscopy. In the ^{13}C NMR mode, the aromatic substrate for oxidation by a dioxygenase is preferably labeled by the ^{13}C isotope. Alkylaryl compounds or the corresponding arylalkanoic acids are synthesized by methods known in the art from a ^{13}C enriched material to incorporate one, or any combination of several, labeled carbon atom(s) into the structure of these compounds. The enrichment levels for any labeled position are preferably at least 5% of ^{13}C, and more preferably at least 95%. Incorporation of ^{13}C label increases sensitivity of the NMR measurement, decreases time required for acquisition of spectrum per sample, and allows for quantitative or semi-quantitative interpretation of compositions of mixtures of isomeric oxidation products. Preferably, incubations with ^{13}C labeled precursors are conducted in multi-well plates, and aliquots of culture fluids or their extracts are sampled with autosampler connected to the solvent line passing through NMR probe without any column separation.

For determining absolute configuration and enantiomeric composition of the HCAs, a variation of the methods described above for determining reaction regioselectivity by ^{1}H or ^{13}C NMR is used. In conjunction with the preferred use of ^{13}C labeled substrates, ^{13}C NMR is preferably employed.

The absolute configuration of any chiral center may be either R or S. In a preferred embodiment, the enantiomeric excess is 98% or more. NMR signals of different enantiomers of HCAs can be distinguished in diastereomeric products using known methods, such as NMR in conjunction with lanthanide shift reagents.

In another preferred embodiment, a variation of the SPA method is used. In this version, a solid support, such as beads or a membrane containing a suitable scintillation dye is used. The solid support is modified with positively charged groups such that it acts

like an anion-exchange material. These materials can be prepared from commercially available SPA materials and they can be used to trap free acids directly in the aqueous medium or culture broths obtained by incubation of the host cells with a radiolabeled alkylarene.

6. Methods for determination of esters

In the interest of brevity, the following discussion focuses on the determination of esters of AHAs. One of skill will appreciate that the same, or similar, methods can be used to determine esters of other compounds formed using the methods of the invention.

Both spectroscopic and non-spectroscopic methods can be used to quantitate the extent of ester synthesis and to characterize the esters. The preferred non-spectroscopic method for assaying AHA methyl ester formation catalyzed by methyl transferases is based on use of a radioactively labeled precursors to AHA methyl esters. ^{14}C or ^{3}H methyl labeled SAM (or its *in-vivo* precursor, methionine) can be used as a probe. In another preferred embodiment, the labeled substrate is the free α-hydroxycarboxylic acid itself.

Using the methods of the invention, methyltransferases that are selective for a particular AHA enantiomer can be selected and further improved by iterative cycles of shuffling and this assay. The selectivity of the methyltransferases of the invention towards a particular enantiomeric configuration of an AHA is preferably measured using samples of the α-hydroxycarboxylic acids that are substantially enantiomerically pure. Host cells employed in this biocatalytic cycle will preferably lack AHA racemase activity (*e.g.* mandelate racemase). In another preferred embodiment, both AHA enantiomers have a different radioactive label, *e.g.* one enantiomer is labeled with ^{14}C, and another with ^{3}H (at one or more H positions which do not readily exchange with water). Measurement of the radioactivity incorporated into the product is performed using a radioactivity detector that allows for the selective measurement of at least two different isotopes. This variation allows the evaluation of the enantioselectivity of a methyltransferases in a single sample.

The radioactivity associated with methyl esters of AHAs is preferably measured in samples which are obtained by selective extraction or partitioning of the methyl esters from neutral or moderately basic (pH about 6-10) aqueous culture samples. These samples can contain varying amounts of free, labeled AHA, of AHA salts and other non-labeled organic compounds. The samples are preferably obtained by incubating individual

clones expressing methyltransferase libraries with the labeled AHAs. The incubation medium is subsequently extracted by adding a defined amount of a preferably water-immiscible organic solvent, or by contacting the broth with an extraction medium (*e.g.* XAD-1180, or similar beads, or membrane).

In those embodiments employing an extraction medium, following its removal from contact with the broth, the extraction media is preferably washed to remove adventitiously bound compounds. Preferred wash solutions are aqueous solutions that do not elute the AHA methyl esters from the extraction medium, but which remove other molecules adsorbed onto the medium. The radioactivity of the extracted material is then measured by methods well known in the art. In embodiments using beads or a membrane an appropriate scintillating dye is preferably used for detecting the radioactivity.

Substantially similar methods can also be employed for detecting other neutral esters of AHAs, such as those exemplified by glycolides (*e.g.*, XVI, **Fig. 13**) and esters of type XX. Thus the same approach is useful for assaying and characterizing the ester forming activity of polypeptides represented by libraries of acyl-transferases, or by a combination of AHA-CoA: alcohol acyltransferases and AHA-CoA ligases. Variations on this method can include the use of a radioactively labeled alcohol (*e.g.*, XIX) or any of its *in-vivo* metabolic precursor.

In another preferred embodiment, the method for detecting polypeptide activity leading to the formation of neutral AHA esters employs UV or fluorescence spectroscopy. This method is applicable to those embodiments in which the transferase activity yields products exhibiting distinct UV and/or fluorescent characteristics. Exemplary compounds include, for example, substituted or non-substituted esters of aromatic carboxylic acids (*e.g.*, mandelic acid). In preferred embodiments of this method, a solvent or solid-phase extraction under neutral or moderately basic conditions (pH about 6-12) is performed on the cell culture medium. Compounds thus isolated are detected by measurement of their UV absorption or fluorescence. These spectral parameters are evaluated to determine relative amounts and identities of the products formed by the transferase reactions.

a. Screening for improved transferase activity

The screening of the transferase libraries, obtained by nucleic acid shuffling or other methods as described above, is done most easily in bacterial or yeast systems by one or more of the screening methods described below.

(i). Methods for detecting increased activity of transferase reactions

The methods for detection of increased formation of monoacyl- and monoglycosyl-derivatives of, for example, glycols and α-hydroxycarboxylic acids include methods in which physical differences between the substrates, the *cis*-diols and the derivatives arising from the transferase-catalyzed reactions are measured. Preferred methods include HPLC and mass-spectrometry. In a high throughput modality, a method of choice is mass-spectrometry, preferably, coordination ion and/or electrospray mass-spectrometry.

For acyl transferases, another presently preferred method uses a labeled acyl-donor precursor, *e.g.* labeled carboxylic acid or its derivative, administered to the cells that express libraries of shuffled genes encoding acyl ligases and/or acyl transferases, *e.g.*, acyl-CoA ligases and acyl-CoA transferases. The amount of label in the hydrophobic reaction products is measured after extraction of the labeled derivatives into a suitable organic solvent, or after solid-phase extraction of these compounds by addition of a sufficient amount of hydrophobic porous resin beads (*e.g.*, XAD 1180, XAD-2, -4, -8). In the case of a radiolabeled compound, scintillating dye can be present in the organic solvent, added to the samples, or chemically incorporated in the bead polymer. The latter constitutes a modification of scintillation proximity assay method.

(ii) Methods for detecting regioselectivity of transferase reactions.

The methods for detecting regioselectivity of the transferase reactions include HPLC, and in an HTP modality, flow-through NMR spectroscopy. When NMR spectroscopy is used for determining relative amounts of different regiomeric monoacyl or monoglycosyl derivatives of oxidized substrates, the latter are preferably obtained by action of the arene dioxygenases on isotopically (^{13}C and/or ^{2}H) labeled substrate. Another variation of the NMR technique includes use of isotopically labeled precursors of acyl- or glycosyl- donor intermediates.

7. Selecting for enhanced organic solvent resistance.

Selection for recombinant polynucleotides that provide improved organic solvent resistance can be accomplished by introducing a library of recombinant polynucleotides into a population of microorganism cells and subjecting the population to a medium that contains various concentrations of the organic hydrophobic compounds of interest. The medium can contain, for example, carbon, nitrogen and minerals, and preferably does not otherwise limit growth and viability of the cells in the absence of the solvent, thus ensuring that solvent resistance is essentially the only limiting factor affecting growth of the cells expressing variants of the genes encoding solvent resistance traits.

In other embodiments, one can employ a screening strategy to identify those recombinant polynucleotides that encode polypeptides that confer improved solvent resistance. For example, one can screen based on the *in vivo* expression of a reporter gene, such as those encoding fluorescent proteins (exemplified by the green fluorescent protein, GFP). Preferably, for the purpose of detecting the best solvent resistant genes under essentially stationary growth phase conditions, those reporter genes are used which display their function in a fashion dependent on availability of intracellular reducing pools, such as NADH and NADPH, and essentially unimpaired ribosomal biosynthesis of proteins.

Such genes and can be exemplified by several bacterial luciferase gene clusters (*lux*) which contain not only luciferase components, but also all polypeptides required for *in-vivo* regeneration of the aldehyde substrate for luciferase.

A variety of methods can be used to detect and to pick or to enrich for the clones with the most efficient solvent resistant traits as judged by display of the properties associated with the *in-vivo* reporter genes. These methods include, for example, fluorescence activating cell sorting of liquid cell suspensions (*e.g.*, cells that express GFP) and CCD camera imaging of individual colonies grown on a solid(ified) medium (*e.g.*, for cells that express *lux*).

If additional improvement in solvent resistance is desired, one can carry out a series of cycles of iterative nucleic acid shuffling and selection by growing the cells in the presence of the organic solvent. Concentrations of the solvents used for selective growth conditions are incrementally increased after each round of recursive mode nucleic acid shuffling in order to provide more stringent selective pressure for those organisms expressing solvent resistance genes.

For use in a high throughput screening protocol, the increase in the solvent resistance to a particular compound of interest and relevance to the biocatalytic synthesis of interest can also be directly measured by administering a radioactively labeled compound and determining relative distribution of radioactivity between cell biomass and extracellular medium components, similar to the method described by Ramos *et al., J. Bacteriol.* **180**:3323-3329 (1998).

G. Bioreactors

In another aspect, the invention provides a bioreactor system for carrying out biotransformations using the improved polypeptides of the invention. The bioreactor includes: (a) an improved dioxygenase polypeptide of the invention; (b) a redox partner source; (c) oxygen; and (d) a substrate for oxidation.

In a preferred embodiment, the dioxygenase polypeptide is an arene dioxygenase polypeptide.

In another preferred embodiment, the bioreactor further includes another useful polypeptide, such as a transferase, ligase, dehydrogenase and the like. The additional useful polypeptide(s) can be co-expressed by a host cell also expressing the improved dioxygenase or expressed by a host cell that does not express the improved dioxygenase. Moreover, each of the polypeptides incorporated into the reactor can be provided as a constituent of a whole cell preparation, a polypeptide extract or as a substantially pure polypeptide. The cells and/or polypeptides are optionally in suspension, in solution, or immobilized on an insoluble matrix, bead or other particle. Additional considerations are discussed below. This discussion is intended as illustrative and not limiting. Other bioreactor formats, conditions, *etc.* will be apparent to those of skill in the art.

General growth conditions for culturing the particular organisms are obtained from depositories and from texts known in the art such as BERGEY'S MANUAL OF SYSTEMATIC BACTERIOLOGY, Vol.1, N. R. Krieg, ed., Williams and Wilkins, Baltimore/London (1984).

For clarity of illustration, the discussion below focuses on the preferred conditions for the oxidation of an organic substrate using the polypeptides of the invention. It is understood that this focus is for the purpose of illustration and that similar conditions are applicable to pathways of the invention other than oxidation.

The nutrient medium for the growth of any oxidizing microorganism should contain sources of assimilable carbon and nitrogen, as well as mineral salts. Suitable sources

of assimilable carbon and nitrogen include, but are not limited to, complex mixtures, such as those constituted by biological products of diverse origin, for example soy bean flour, cotton seed flour, lentil flour, pea flour, soluble and insoluble vegetable proteins, corn steep liquor, yeast extract, peptones and meat extracts. Additional sources of nitrogen are ammonium salts and nitrates, such as ammonium chloride, ammonium sulfate, sodium nitrate and potassium nitrate. Generally, the nutrient medium should include, but is not limited to, the following ions: Mg^{2+}, Na^{+}, K^{+}, Ca^{2+}, NH_4^{+}, Cl^{-}, SO_4^{2-}, PO_4^{2-} and NO_3^{-} and also ions of the trace elements such as Cu, Fe, Mn, Mo, Zn, Co and Ni. The preferred source of these ions are mineral salts.

If these salts and trace elements are not present in sufficient amounts in the complex constituents of the nutrient medium or in the water used it is appropriate to supplement the nutrient medium accordingly.

The microorganisms employed in the process of the invention can be in the form of fermentation broths, whole washed cells, concentrated cell suspensions, polypeptide extracts, and immobilized polypeptides and/or cells. Preferably concentrated cell suspensions, *polypeptide* extracts, and whole washed cells are used with the process of the invention (S. A. White and G. W. Claus, *J. Bacteriology,* **150**:934-943 (1982)).

Methods of immobilizing polypeptides and cells are well known in the art and include such techniques as microencapsulation, attachment to alginate beads, cross-linked polyurethane, starch particles, polyacrylamide gels and the use of coacervates, which are aggregates of colloidal droplets. In a presently preferred embodiment, the polypeptide and/or cell is immobilized onto a glass particles having a porous outer surface, such as that described in Dubin , *et al.*, U.S. Patent No. 5,922,531, issued July 13, 1999.

Concentrated washed cell suspensions may be prepared as follows: the microorganisms are cultured in a suitable nutrient solution, harvested (for example by centrifuging) and suspended in a smaller volume (in salt or buffer solutions, such as physiological sodium chloride solution or aqueous solutions of potassium phosphate, sodium acetate, sodium maleate, magnesium sulfate, or simply in tap water, distilled water or nutrient solutions). The substrate is then added to a cell suspension of this type and the oxidation reaction according to the invention is carried out under the conditions described.

The conditions for oxidizing a substrate in growing microorganism cultures or fractionated cell extracts are advantageous for carrying out the process according to the invention with concentrated cell suspensions. In particular the temperature range is from

about 0 °C to about 45 °C and the pH range is from about 2 to about 10. There are no special nutrients necessary in the process of the invention. More importantly, washed or immobilized cells can simply be added to a solution of substrate, without any nutrient medium present.

It is also possible to carry out the process according to the invention with polypeptide extracts or polypeptide extract fractions prepared from cells. The extracts can be crude extracts, such as obtained by conventional digestion of microorganism cells. Methods to break up cells include, but are not limited to, mechanical disruption, physical disruption, chemical disruption, and enzymatic disruption. Such means to break up cells include ultrasonic treatments, passages through French pressure cells, grindings with quartz sand, autolysis, heating, osmotic shock, alkali treatment, detergents, or repeated freezing and thawing.

If the process according to the invention is to be carried out with partially purified polypeptide extract preparations, the methods of protein chemistry, such as ultracentrifuging, precipitation reactions, ion exchange chromatography or adsorption chromatography, gel filtration or electrophoretic methods, can be employed to obtain such preparations. In order to carry out the reaction according to the invention with fractionated cell extracts, it may be necessary to add to the assay system additional reactants such as, physiological or synthetic electron acceptors, like NAD^+, $NADP^+$, methylene blue, dichlorophenolindophenol, tetrazolium salts and the like. When these reactants are used, they can be employed either in equimolar amounts (concentrations which correspond to that of the substrate employed) or in catalytic amounts (concentrations which are markedly below the chosen concentration of substrate). If, when using catalytic amounts, it is to be ensured that the process according to the invention is carried out approximately quantitatively, a system which continuously regenerates the reactant which is present only in a catalytic amount must also be added to the reaction mixture. This system can be, for example, a polypeptide which ensures reoxidation (in the presence of oxygen or other oxidizing agents) of an electron acceptor which is reduced in the course of the reaction according to the invention.

If nutrient media is used with intact microorganisms in a growing culture, nutrient media can be solid, semi-solid or liquid. Aqueous-liquid nutrient media are preferably employed when media is used. Suitable media and suitable conditions for cultivation include known media and known conditions to which substrate can be added.

The substrate to be oxidized in the process of the invention can be added to the base nutrient medium either on its own or as a mixture with one or more oxidizable compounds. Additional oxidizable compounds which can be used include polyols, such as sorbitol or glycerol.

If one or more oxidizable compounds are added to the nutrient solution, the substrate to be oxidized can be added either prior to inoculation or at any desired subsequent time (between the early log phase and the late stationary growth phase). In such a case the oxidizing organism is preferably pre-cultured with the oxidizable compounds. The inoculation of the nutrient media is effected by a variety of methods including slanted tube cultures and flask cultures.

Contamination of the reaction solution should be avoided. To avoid contamination, sterilization of the nutrient media, sterilization of the reaction vessels and sterilization of the air required for aeration is preferably undertaken. It is possible to use, for example, steam sterilization or dry sterilization for sterilization of the reaction vessels. The air and the nutrient media can likewise be sterilized by steam or by filtration. Heat sterilization of the reaction solution containing the substrate is also possible.

The process of the invention can be carried out under aerobic conditions using shake flasks or aerated and agitated tanks. Preferably, the process is carried out by the aerobic submersion procedure in tanks, for example in conventional fermentors. It is possible to carry out the process continuously or with batch or fed batch modes, preferably the batch mode.

It is advantageous to ensure that the microorganisms are adequately brought into contact with oxygen and the substrate. This can be effected by several methods including shaking, stirring and aerating.

If foam occurs in an undesired amount during the process, chemical foam control agents, such as liquid fats and oils, oil-in-water emulsions, paraffins, higher alcohols (such as octadecanol), silicone oils, polyoxyethylene compounds and polyoxypropylene compounds, can be added. Foam can also be suppressed or eliminated with the aid of mechanical devices.

H. Example of Dioxygenase Shuffling

The toluene dioxygenase genes *tod*C1C2BA (Zylstra G. J. and D. T. Gibson. (1989) *J. Biol. Chem.* 264: 14940-14946) were used as a substrate for shuffling a family of genes.

The length of this four gene operon is 3.591 kb. The tetrachlorobenzene dioxygenase genes *tec*A1A2A3A4 (Biel *et al.* (1997) *Eur. J. Biochem.* 247:190-199) were also used as a substrate for shuffling a family of genes. The length of this four gene operon is 3.587 kb. After shuffling the family of genes, the shuffled sequences were cloned into the expression vector, pTrc99a (Pharmacia) and transformed into *E. coli* cells. The transformed cells were plated onto LB (Luria Bertani) agar plates containing ampicillin and incubated at 37°C for 20 hr. Restriction digests were used to determine that 23 out of the 24 randomly selected library clones contained sequences that were chimeras of the two parental sequences.

TIER 1 ASSAY

The library was screened for active clones using a plate assay which detects the oxidation of indole to indigo by an active dioxygenase enzyme (Ensley *et al.* (1983) *Science* 222: 167-169). The library was plated onto LB agar plates containing ampicillin and incubated at 37°C for 20 hr after which the plates are placed at 23°C and incubated for an additional 24-48 hr. Clones expressing active dioxygenases produced colored colonies from accumulation of indigo. The color ranged from blue to blue-grey and the intensity of the color varied from clone to clone depending on the level of enzyme activity. The library was determined to be 70% active. The colored colonies were picked using an Q-bot (Genetix) into 384-well plates containing LB ampicillin and incubated at 37°C with shaking for 20 hr. The library plates were stored after the addition of sterile glycerol (10% final concentration) at –20°C until further screening.

TIER II ASSAY

The following high-throughput (HTP) assay was developed and used to identify shuffled dioxygenases with improved oxidation of *p*-xylene to *p*-xylene diol. The assay is not limited to the substrate, *p*-xylene, and can be used for any volatile or toxic (to the test organism) substrate.

Library clones are grown to saturation in a 96-well plate(s) with 250ul 2xYT containing ampicillin and 0.2%glucose. The plate(s) are inoculated with 3 ul/well inoculum. The plate(s) are incubated at 250 rpm, 37°C, with 85-90% humidity for 20 hr. The cultures are subcultured into deep 96-well induction plate(s) containing 0.5 ml 2xYT containing ampicillin and 1mM IPTG/well. The plate(s) are incubated at 37°C, 250 rpm, with 85-90% humidity, 6 hr. The induced cells are harvested by centrifugation at 3000 rpm, 10 min., 10°C. The cells are washed once with minimal media and resuspended to a final volume of 0.4 ml. The cells are transferred to an assay plate(s)

containing the volatile or toxic substrate to be tested embedded in solidified 1% agarose on the bottom of each well of the plate(s). The assay plate(s) are sealed and incubated at 23°C for 1 hr with vigorous shaking. The cell suspension is transferred to a clean polypropylene 96-well plate(s) and the cells pelleted by centrifugation at 3000 rpm, 10 min, at 10°C. The cell-free supernatants are transferred to a 96-well plate(s) and the diol product analyzed by spectrophotometric methods, HPLC, or GC-MS. Potential positive clones are those clones with improved properties when compared to the best parent.

Using the HTP assay described above, 1080 library clones were screened and compared to the parental dioxygenases for the ability to oxidize *p*-xylene to *p*-xylene diol. Nine clones were identified as potential positives and retested to confirm the improvements. See table 1 for levels of improvement and Figure 16 to illustrate crossovers in these nine clones. The nine clones have 2.2 to 3.7-fold improved activity when compared to the best parent. In addition, the two parent sequences and the nine potential positive clones were transformed into an alternative *E. coli* host strain and tested for oxidation of *p*-xylene to *p*-xylene diol using the HTP plate assay. The nine clones and the parents were tested after growth in two types of media; 2xYT and minimal media with fructose as the carbon source. Of the nine clones, six were reconfirmed as 1.7-2.8-fold improved over the best parent.

Positives	1st Screen	-fold improv.	2nd screen		-fold improv.	ave. read	(stdev)	ave. improv.
Clone 1	2.99	1.6	3.85	3.17	2.5,2.6	3.336667	0.453578	2.2+/-.44
Clone 2	3.25	1.7	3.89	3.07	2.5,2.6	3.403333	0.430968	2.3+/-.40
Clone 3	6.53	3.5 +	5.14	4.97	3.4,4.2	5.546667	0.855823	3.7+/-.35
Clone 4	3.47	2.6 +	4.15	4.78	2.6,4.2	4.133333	0.655159	3.1+/-.75
Clone 5	4.22	2.5 +	3.84	3.49	2.5,2.9	3.85	0.365103	2.6+/-.18
Clone 6	5.2	3.8	4.58	3.64	2.8,3.2	4.473333	0.785451	3.3+/-.41
Clone 7	4.14	3.6 +	3.84	3.01	2.5,2.6	3.663333	0.585349	2.9+/-.49
Clone 8	2.47	1.6 +	4.05	4.22	2.5,3.7	3.58	0.965039	2.6+/-.86
Clone 9	5.01	3.6 +	4.05	3.75	2.5,3.3	4.27	0.658179	3.1+/-.46

I. Kits

Also provided is a kit or system utilizing any one of the selection strategies, materials, components, methods or substrates herein before described. Kits will optionally additionally include instructions for performing methods or assays, packaging materials, one or more containers which contain assay, device or system components, or the like.

In an additional aspect, the present invention provides kits embodying the methods and apparatus herein. Kits of the invention optionally include one or more of the following: (1) a shuffled component as described herein; (2) instructions for practicing the methods described herein, and/or for operating the selection procedure herein; (3) one or more dioxygenase assay component; (4) a container for holding dioxygenase nucleic acids or polypeptides, other nucleic acids, transgenic plants, animals, cells, or the like and, (5) packaging materials.

In a further embodiment, the present invention provides for the use of any composition or kit herein, for the practice of any method or assay herein, and/or for the use of any apparatus or kit to practice any assay or method herein.

In yet another embodiment, the kit of the invention includes one or more improved dioxygenase polypeptides of the invention. In a preferred embodiment, the kit includes a library of improved dioxygenase polypeptides.

While the foregoing invention has been described in some detail for purposes of clarity and understanding, it will be clear to one skilled in the art from a reading of this disclosure that various changes in form and detail can be made without departing from the true scope of the invention. For example, all the techniques and apparatus described above may be used in various combinations. All publications, patent applications, patents, and other documents cited in this application are incorporated by reference in their entirety for all purposes to the same extent as if each individual publication or patent document were individually so denoted.

WHAT IS CLAIMED IS:

1. A method for obtaining a nucleic acid encoding an improved polypeptide comprising dioxygenase activity, wherein said improved polypeptide has at least one property improved over a naturally occurring dioxygenase polypeptide, said method comprising:

(a) creating a library of recombinant polynucleotides encoding one or more recombinant dioxygenase polypeptides; and

(b) screening the library to identify a recombinant polynucleotide encoding an improved recombinant dioxygenase polypeptide that has at least one property improved over said naturally occurring polypeptide.

2. The method according to claim **1**, wherein said creating a library comprises:

shuffling a plurality of parental polynucleotides to produce one or more recombinant dioxygenase polynucleotide encoding said improved recombinant dioxygenase polypeptide.

3. The method according to claim **1**, wherein said improved polypeptide converts to a vicinal diol at least one π-bond selected from an olefinic π-bond, an aryl π-bond and a heteroaryl π-bond.

4. The method of claim **2**, wherein at least one of said parental polynucleotides encodes a polypeptide having at least one dioxygenase activity.

5. The method of claim **2**, wherein said parental polynucleotides are homologous.

6. The method of claim **2**, wherein at least one of said parental polynucleotides does not encode a dioxygenase activity.

7. The method of claim **4**, wherein said parental dioxygenase polynucleotide encodes a polypeptide or polypeptide subsequence selected from arene dioxygenase polypeptides.

8. The method of claim **2**, wherein a member selected from said parental polynucleotides, said one or more recombinant dioxygenase polynucleotide, said

identified recombinant dioxygenase polynucleotide and combinations thereof is cloned into an expression vector.

9. The method of claim **1**, wherein said identified recombinant dioxygenase polynucleotide catalyzes an enzymatic reaction using a redox partner other than NADH.

10. The method of claim **2**, further comprising:

creating a library of recombinant dehydrogenase activity polynucleotides encoding a recombinant dehydrogenase activity;

screening said library to identify a recombinant polynucleotide that encodes an improved dehydrogenase activity; and

co-expressing one or more of said identified recombinant dehydrogenase activity polynucleotides and said identified recombinant dioxygenase polynucleotide in a cell.

11. The method of claim **1**, further comprising:

creating a library of recombinant transferase activity polynucleotides encoding a recombinant transferase activity;

screening said library to identify a recombinant polynucleotide that encodes an improved transferase activity; and

co-expressing one or more of said identified recombinant transferase activity polynucleotides and said identified recombinant dioxygenase polynucleotide in a cell.

12. The method according to claim **11**, wherein said transferase polynucleotide encodes a polypeptide that is a member selected from acyltransferases, glycosyltransferases, methyl transferases and combinations thereof.

13. The method of claim **2**, wherein said plurality of parental polynucleotides are shuffled to produce a library of recombinant polynucleotides comprising one or more library member polynucleotide encoding one or more dioxygenase activity, which library is selected for one or more dioxygenase activity, selected from converting a π-bond to a vicinal diol.

14. The method of claim **13**, wherein said π-bond is selected from an olefinic π-bond, an aryl π-bond, a heteroaryl π-bond and combinations thereof.

15. A library of recombinant polynucleotides encoding one or more dioxygenase activity made by the method of claim **1**.

16. The library of claim **15**, wherein said library is a phage display library.

17. An improved dioxygenase encoding nucleic acid prepared by the method according to claim **1**.

18. A library of improved dioxygenase encoding nucleic acids, said library comprising at least two dioxygenase encoding nucleic acids prepared by the method according to claim **1**.

19. The method of claim **2**, wherein said parental polynucleotides are shuffled in a plurality of cells selected from prokaryotes and eukaryotes.

20. The method of claim **2**, wherein said parental polynucleotides are shuffled in a plurality of cells selected from bacteria, yeast, and fungi.

21. The method of claim **2**, wherein said parental polynucleotides are shuffled in a plurality of cells, said method further comprising one or more members selected from:

(a) recombining DNA from said plurality of cells that display dioxygenase activity with a library of DNA fragments, at least one of which undergoes recombination with a segment in a cellular DNA present in said cells to produce recombined cells, or recombining DNA between said plurality of cells that display dioxygenase activity to produce cells with modified dioxygenase activity;

(b) recombining and screening said recombined or modified cells to produce further recombined cells that have evolved additionally modified dioxygenase activity; and

(c) repeating (a) or (b) until said further recombined cells have acquired a desired dioxygenase activity.

22. The method of claim **2**, wherein said method further comprises:

(a) recombining at least one distinct or improved recombinant polynucleotide with a further dioxygenase activity polynucleotide, which further polynucleotide is identical to or different from one or more of said plurality of parental polynucleotides to produce a library of recombinant dioxygenase polynucleotides;

(b) screening said library to identify at least one further distinct or improved recombinant dioxygenase polynucleotide that exhibits a further improvement or distinct property compared to said plurality of parental polynucleotides; and, optionally,

(c) repeating (a) and (b) until said resulting further distinct or improved recombinant polynucleotide shows an additionally distinct or improved dioxygenase property.

23. The method of claim **2**, wherein said recombinant dioxygenase polynucleotide is present in one or more cells selected from bacterial, yeast and fungal cells and said method comprises:

pooling multiple separate dioxygenase polynucleotides;

screening said resulting pooled dioxygenase polynucleotides to identify an improved recombinant dioxygenase polynucleotides that exhibits an improved dioxygenase activity compared to a non-recombinant dioxygenase activity polynucleotide; and

cloning said improved recombinant nucleic acid.

24. The method of claim **23**, further comprising transducing said distinct or improved nucleic acid into a member selected from a prokaryote and a eukaryote.

25. The method of claim **2**, wherein said shuffling of a plurality of parental polynucleotides comprises family gene shuffling.

26. The method of claim **2**, wherein said shuffling of a plurality of parental nucleic acids comprises individual gene shuffling.

27. A selected shuffled dioxygenase nucleic acid made by the method of claim **2**.

28. A nucleic acid shuffling mixture, comprising: at least three homologous DNAs, each of which is derived from a polynucleotide encoding a member

selected from a polypeptide encoding dioxygenase activity, a polypeptide fragment encoding dioxygenase activity and combinations thereof.

29. The nucleic acid shuffling mixture of claim **28**, wherein said at least three homologous DNAs are present in cell culture or *in vitro*.

30. A method for increasing dioxygenase activity in a cell, comprising: performing whole genome shuffling of a plurality of genomic polynucleotides in said cell and selecting for one or more dioxygenase activity.

31. The method of claim **30**, wherein said genomic polynucleotides are from a species or strain different from said cell.

32. The method of claim **30**, wherein said cell is of prokaryotic or eukaryotic origin.

33. The method of claim **30**, wherein said dioxygenase activity to be selected comprises converting to a vicinal diol at least one π-bond selected from an olefinic π-bond, an aryl π-bond, a heteroaryl π-bond and combinations thereof.

34. A method for obtaining a polynucleotide encoding an improved dioxygenase polypeptide acting on a substrate comprising a target group selected from an olefin, an aryl group and combinations thereof, wherein said improved polypeptide exhibits one or more improved properties compared to a naturally occurring polypeptide acting on said substrate, said method comprising:

(a) creating a library of recombinant polynucleotides encoding a dioxygenase polypeptide acting on said substrate; and

(b) screening said library to identify a recombinant polynucleotide encoding an improved polypeptide that exhibits one or more improved properties compared to a naturally occurring dioxygenase polypeptide.

35. The method according to claim **34**, wherein said library of recombinant polynucleotides is created by recombining at least a first form and a second form of a nucleic acid, at least one form encoding said naturally occurring polypeptide or a fragment thereof, wherein said first form and said second form differ from each other in two or more nucleotides.

36. The method according to claim **35**, wherein said first and second forms of said nucleic acid are homologous.

37. The method according to claim **35**, wherein at least one of said first and second forms of said nucleic acid does not encode a polypeptide having dioxygenase activity.

38. A dioxygenase polypeptide encoded by a nucleic acid according to claim **1**.

39. A dioxygenase polypeptide encoded by the polynucleotide according to claim **34**.

40. A library of two or more dioxygenase polypeptides, each of said polypeptides being encoded by a nucleic acid according to claim **1**.

41. A library of two or more dioxygenase polypeptides, each of said polypeptides being encoded by a nucleic acid according to claim **34**.

42. The polypeptide according to claim **39** wherein said polypeptide has an activity comprising, converting an olefin to a diol.

43. The polypeptide according to claim **39**, wherein said polypeptide has an activity comprising, converting at least one π-bond of an aryl group to a diol.

44. The polypeptide according to claim **39**, wherein said improved property is selected from:

improved regiospecificity of said acting on a substrate, wherein said substrate comprises at least two target groups;

enhanced production of a desired enantiomeric form of a reaction product;

enhanced expression of said polypeptide by a host cell that comprises said recombinant polynucleotide; and

enhanced stability of said polypeptide in said presence of an organic solvent.

45. A method of oxidizing a substrate comprising a target group selected from an olefin, an aryl group, a heteroaryl group and combinations thereof, said method comprising contacting said substrate with a polypeptide according to claim **39**

46. The method according to claim **45**, wherein said absolute configuration of a product of said dioxygenase is R, S, or a mixture thereof.

47. A method for preparing a diol group, said method comprising contacting a substrate comprising a π-bond with a polypeptide according to claim **42**.

48. The method according to claim **47**, in which said π-bond is a carbon-carbon bond.

49. A method for preparing a hydroxyaryl group, said method comprising contacting a substrate comprising an aryl group with a polypeptide according to claim **43**.

50. An organism comprising a recombinant dioxygenase polynucleotide encoding an improved dioxygenase polypeptide that catalyzes a reaction selected from converting to a vicinal diol at least one π-bond selected from an olefinic π-bond, an aryl π-bond, a heteroaryl π-bond and combinations thereof, wherein said polypeptide exhibits one property improved relative to a corresponding property of a naturally occurring dioxygenase polypeptide.

51. The organism according to claim **50**, further comprising a transferase polypeptide.

52. The organism according to claim **51**, wherein said transferase polypeptide is an improved transferase polypeptide that exhibits one or more properties improved relative to a corresponding property of a naturally occurring transferase.

53. The organism according to claim **51**, wherein said transferase is selected from S-adenosylmethionine dependent O-methyltransferase, acyl-CoA transferase and combinations thereof.

54. The organism according to claim **50**, further comprising a ligase polypeptide.

55. The organism according to claim **54**, wherein said ligase polypeptide is an improved ligase polypeptide that exhibits one or more properties improved relative to a corresponding property of a naturally occurring ligase.

56. The organism according to claim **54**, wherein said ligase is an acyl CoA ligase.

57. The organism according to claim **50**, further comprising a racemase polypeptide.

58. The organism according to claim **57**, wherein said racemase polypeptide is an improved racemase polypeptide that exhibits one or more properties improved relative to a corresponding property of a naturally occurring transferase.

59. The organism according to claim **57**, wherein said racemase is mandelate racemase.

60. The organism according to claim **50**, further comprising a dehydrogenase polypeptide.

61. The organism according to claim **60**, wherein said dehydrogenase polypeptide is an improved dehydrogenase polypeptide that exhibits one or more improved properties improved relative to a corresponding property of a naturally occurring dehydrogenase.

62. The organism according to claim **50**, further comprising a solvent resistance polypeptide that confers upon said organism a resistance to an organic solvent.

63. The organism according to claim **62**, wherein said solvent resistance polypeptide is an improved solvent resistance polypeptide that exhibits one or more improved properties improved relative to a corresponding property of a naturally occurring solvent resistance polypeptide.

64. The organism according to claim **62**, wherein said improved solvent resistance polypeptide imparts to the organism a resistance to one or more organic compounds selected from olefins, α-hydroxycarboxylic acids, diols, aldehydes, ketones, halogenated hydrocarbons, perfluorocarbons, esters, aryl compounds, carboxylic acids, alcohols, ethers and combinations thereof.

65. The organism of claim **62**, wherein said improved solvent resistance polypeptide imparts to the organism a resistance to said solvent, wherein the solvent is present in a medium at hypersaturating concentrations.

66. The organism of claim **50**, wherein said organism further comprises two or more recombinant polynucleotides selected from the group consisting of:

an improved transferase polypeptide that exhibits one or more properties improved relative to a corresponding property of a naturally occurring transferase polypeptide;

an improved ligase peptide that exhibits one or more properties improved relative to a corresponding property of a naturally occurring ligase polypeptide;

an improved racemase polypeptide that exhibits one or more properties improved relative to a corresponding property of a naturally occurring racemase polypeptide;

an improved dehydrogenase polypeptide that exhibits one or more properties improved relative to a corresponding property of a naturally occurring dehydrogenase polypeptide;

an improved solvent resistance polypeptide that confers upon said organism a resistance to an organic solvent that is improved relative to that conferred by a naturally occurring solvent resistance-conferring polypeptide.

67. A method for preparing a vicinal diol group, said method comprising:

(a) contacting a substrate comprising a carbon-carbon double bond with an organism according to claim **50**, thereby forming said vicinal diol group.

68. The method according to claim **67**, wherein said substrate is selected from α-olefins and n-alkenes.

69. The method according to claim **67**, wherein said substrate is selected from styrene, substituted styrene, divinylbenzene, substituted divinylbenzene, isoprene, butadiene, diallyl ether, allyl phenyl ether, substituted allyl phenyl ether, allyl alkyl ether, allyl aralkyl ether, vinylcyclohexene, vinylnorbornene, and acrolein.

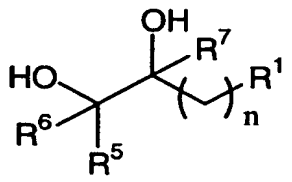
70. The method according to claim **67**, further comprising,

(b) contacting said vicinal diol with an organism comprising a polypeptide selected from a ligase polypeptide, a transferase polypeptide and combinations thereof, thereby forming a vicinal diol adduct.

71. The method according to claim **70**, wherein said polypeptide of (b) is a polypeptide exhibiting one or more properties improved over a corresponding property of an analogous naturally occurring polypeptide.

72. The method according to claim **70**, wherein said polypeptide of (a), and said polypeptide of (b) are expressed in the same host.

73. The method according to claim **67**, wherein said vicinal diol has the structure:



wherein

R^1 and R^5 are independently selected from alkyl, substituted alkyl, aryl, substituted aryl, heteroaryl, substituted heteroaryl, heterocyclic, substituted heterocyclic, $—NR^2R^3$, $—OR^2$, —CN, $C(R^4)NR^2R^3$ and $C(R^4)OR^2$ groups, or R^1 and R^5 are joined to form a ring system selected from saturated hydrocarbyl rings, unsaturated hydrocarbyl rings, saturated heterocyclic rings and unsaturated heterocyclic rings;

R^2 and R^3 are independently selected from H, alkyl, substituted alkyl, aryl, substituted aryl, heteroaryl, substituted heteroaryl, heterocyclic and substituted heterocyclic groups;

R^4 is selected from =O and =S;

R^6 and R^7 are independently selected from H and alkyl; and

n is an integer from 0 to 10, inclusive.

74. The method according to claim **73**, wherein

R^1 is selected from phenyl, substituted phenyl, pyridyl, substituted pyridyl $—NR^2R^3$, $—OR^2$, —CN, $C(R^4)NR^2R^3$ and $C(R^4)OR^2$ groups,

R^2 and R^3 are independently selected from H, alkyl, substituted alkyl, aryl, substituted aryl, heteroaryl, substituted heteroaryl, heterocyclic and substituted heterocyclic groups; and

R^4 is selected from =O and =S.

75. The method according to claim **74**, wherein said diol comprises a six-member ring having at least one intraannular double bond and at least one substituent selected from methyl, carboxyl and combinations thereof.

76. The method according to claim **75**, wherein said diol is a member selected from compounds III, IV, V, VI, VII, VIII, XXIII, XXIV, XXV, XXVI, XXVII, XXVIII, XXIX, and XXX.

77. A method for converting an olefin into an α-hydroxyacid, said method comprising:

(a) contacting said olefin with an organism according to claim **50** to form a vicinal diol; and

(b) contacting said vicinal diol with an organism comprising a dehydrogenase polypeptide to form said α-hydroxyacid.

78. The method according to claim **77**, wherein said dehydrogenase polypeptide exhibits at least one property improved relative to a corresponding property in an analogous naturally occurring polypeptide.

79. The method according to claim **77**, wherein said polypeptide of (a), and of (b) are expressed in the same host.

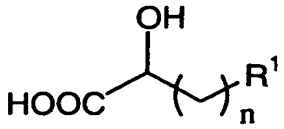
80. The method according to claim **77**, further comprising:

(c) contacting said α-hydroxyacid with an organism comprising an improved polypeptide having an activity selected from ligase, transferase and combinations thereof, thereby forming a α-hydroxyacid adduct.

81. The method according to claim **80**, wherein at least two of said polypeptides of (a), (b), and (c) are expressed in the same host.

82. The method according to claim **80**, wherein at least one of said polypeptides selected from ligase, transferase and combinations thereof is an improved polypeptide.

83. The method according to claim **77**, wherein said α-hydroxycarboxylic acid has the structure:



wherein

R^1 is selected from aryl, substituted aryl, heteroaryl, substituted heteroaryl, heterocyclic, substituted heterocyclic, —NR^2R^3, —OR^2, —CN, $C(R^4)NR^2R^3$ and $C(R^4)OR^2$ groups,

R^2 and R^3 are independently selected from H, alkyl, substituted alkyl, aryl, substituted aryl, heteroaryl, substituted heteroaryl, heterocyclic and substituted heterocyclic groups;

R^4 is selected from =O and =S, and

n is an integer between 0 and 10, inclusive.

84. The method according to claim **83** wherein

R^1 is selected from phenyl, substituted phenyl, pyridyl, substituted pyridyl, —NR^2R^3, —OR^2, —CN, $C(R^4)NR^2R^3$ and $C(R^4)OR^2$ groups,

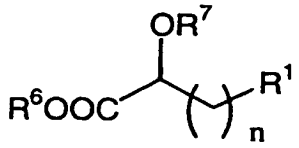
R^2 and R^3 are independently selected from H, alkyl, substituted alkyl, aryl, substituted aryl, heteroaryl, substituted heteroaryl, heterocyclic and substituted heterocyclic groups; and

R^4 is selected from =O and =S.

85. The method according to claim **80**, wherein said transferase activity is selected from glycosyl transferase activity and methyltransferase activity.

86. The method according to claim **85**, wherein said methyl transferase is a S-adenosylmethionine dependent O-methyltransferase.

87. The method according to claim **80**, wherein said α-hydroxyacid adduct has the structure:



wherein

R^1 is selected from aryl, substituted aryl, heteroaryl, substituted heteroaryl, heterocyclic, substituted heterocyclic, —$NR^2R^3(R^4)_m$, —OR^2, —CN, $C(R^5)NR^2R^3$ and $C(R^5)OR^2$ groups,

R^2, R^3 and R^4 are members independently selected from said group consisting of H, alkyl, substituted alkyl, aryl, substituted aryl, heteroaryl, substituted heteroaryl, heterocyclic and substituted heterocyclic groups;

R^5 is selected from =O and =S;

R^6 is selected from H, alkyl and substituted alkyl groups;

R^7 is $C(O)R^8$, wherein R^8 is selected from H, alkyl, and substituted alkyl groups and wherein R^7 and R^8 are not both H;

m is 0 or 1, such that when m is 1, an ammonium salt is provided; and

n is an integer between 0 and 10, inclusive.

88. The method according to claim **87** wherein

R^1 is selected from phenyl, substituted phenyl, pyridyl, substituted pyridyl —NR^2R^3, —OR^2, —CN, $C(R^5)NR^2R^3$ and $C(R^5)OR^2$ groups

R^2 and R^3 are independently selected from said group consisting of H, C_1-C_6 alkyl and allyl; and

R^5 is =O.

89. A method for preparing a hydroxylated aromatic carboxylic acid comprising:

(a) contacting a substrate comprising an aryl carboxylic acid with an organism according to claim **50**.

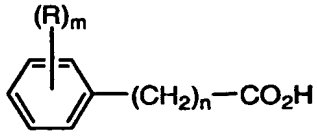
90. A method according to claim **89**, wherein said aryl carboxylic acid is produced by contacting a substrate comprising an aryl alkyl group with a monooxygenase polypeptide, thereby producing said aryl carboxylic acid.

91. The method according to claim **104**, wherein said monooxygenase polypeptide and said dioxygenase polypeptide are expressed in the same host.

92. The method according to claim **90**, wherein said monooxygenase polypeptide is an improved polypeptide exhibiting at least one property improved relative to a corresponding property of a naturally occurring monooxygenase polypeptide.

93. The method according to claim **89**, wherein said aryl group is selected from substituted aryl and substituted heteroaryl groups.

94. The method according to claim **93**, wherein said substituted aryl group has the structure:



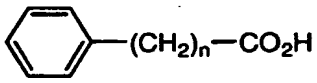
wherein

each of said m R groups is selected from alkyl, substituted alkyl, aryl; substituted aryl, heteroaryl, substituted heteroaryl, heterocyclic and substituted heterocyclic;

m is an integer from 0 to 5, inclusive; and

n is an integer from 1 to 10, inclusive.

95. The method according to claim **81**, wherein said substituted aryl group has the structure:

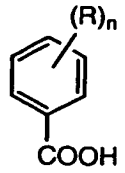


wherein

n is an integer from 1 to 6, inclusive.

96. The method according to claim **93**, wherein said substituted aryl group is substituted with at least one methyl moiety.

97. The method according to claim **89**, wherein said aryl group is



wherein

each of said n R groups is independently selected from H, alkyl and substituted alkyl groups; and

n is an integer from 1 to 5, inclusive.

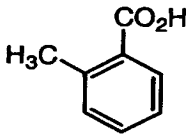
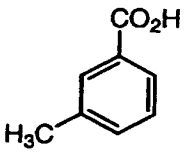
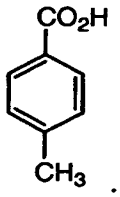
98. The method according to claim **106**, wherein

R is methyl; and

n is an integer from 1 to 5, inclusive.

99. The method according to claim **98**, wherein n is an integer from 1 to 3, inclusive.

100. The method according to claim **99**, wherein said aryl group is selected from:

 ,  and  .

101. The method according to claim **89**, further comprising:

(b) contacting said hydroxylated aromatic carboxylic acid with an organism comprising an improved polypeptide having an activity selected from ligase, transferase and combinations thereof, thereby forming a hydroxylated aromatic carboxylic acid adduct.

102. The method according to claim **80**, wherein said polypeptides of (a), and (b) are expressed in the same host.

103. The method according to claim **80**, wherein at least one of said polypeptides selected from ligase, transferase and combinations thereof is an improved polypeptide.

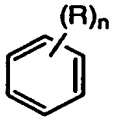
104. A method for preparing a hydroxyaryl group, said method comprising:

(a) contacting a substrate comprising aryl group with a microorganism according to claim **50**, to form a vicinal diol;

(b) dehydrating said vicinal diol, thereby forming said hydroxyaryl group.

105. The method according to claim **104**, wherein said substrate comprises a member selected from arylalkyl groups, substituted arylalkyl groups, heteroarylalkyl groups, and substituted heteroarylalkyl groups.

106. The method according to claim **104**, wherein said substrate has the structure



wherein,

each of said n R groups is a member selected from the group consisting of alkyl groups and substituted alkyl groups; and

n is an integer from 0 to 5, inclusive.

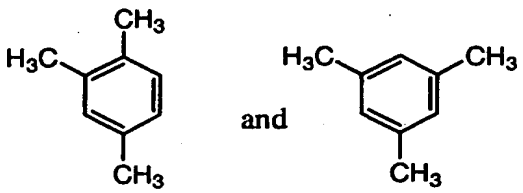
107. The method according to claim **106**, wherein

R is methyl; and

n is an integer from 1 to 5, inclusive.

108. The method according to claim **107**, wherein n is an integer from 1 to 3, inclusive.

109. The method according to claim **108**, wherein said aryl group is selected from:



110. The method according to claim **104**, further comprising:

prior to (b), contacting said vicinal diol with an organism comprising a polypeptide having transferase activity, thereby forming a vicinal diol adduct.

111. The method according to claim **110**, wherein said transferase activity is selected from acyltransferase activity, glycosyltransferase activity and combinations thereof.

112. The method according to claim **110**, wherein said contacting with a polypeptide having transferase activity occurs prior to said dehydrating said vicinal diol.

113. The method according to claim **110**, wherein said polypeptides of (a), and (b) are expressed in the same host.

114. The method according to claim **110**, wherein said transferase exhibits one or more properties improved relative to a corresponding property of a naturally occurring glycosyltransferase.

115. A bioreactor comprising:

(a) an improved dioxygenase polypeptide;

(b) a redox partner;

(c) oxygen; and

(d) an oxidizable substrate.

116. The bioreactor according to claim **115**, wherein said polypeptide is immobilized.

117. The method of claim **2**, wherein said shuffling comprises:

(1) recombining at least first and second forms of a nucleic acid that encodes a dioxygenase polypeptide, or fragment thereof, wherein the first and second forms differ from each other in two or more nucleotides, to produce the library of recombinant polynucleotides; and

(2) expressing the library of recombinant polynucleotides to obtain a library of recombinant polypeptides.

118. The method of claim **117**, wherein step (1) is performed *in vitro*.

119. The method of claim **117**, wherein the method further comprises:

(3) recombining at least one recombinant polynucleotide that encodes a member of the library of recombinant polynucleotides that encodes a member of the library of recombinant dioxygenase polypeptides, which is the same or different from the first and second forms, to produce a further library of recombinant polynucleotides;

(4) expressing the further library of recombinant polynucleotides to obtain a further library of recombinant dioxygenase polypeptides; and

(5) repeating (3) and (4), as necessary, until the further library of recombinant polynucleotides contains a desired number of different recombinant polynucleotides.

120. The method of claim **119**, wherein at least one recombining step is performed *in vitro*.

121. The method of claim **2**, wherein the shuffling comprises:

(1) hybridizing at least two sets of nucleic acids, wherein a first set of nucleic acids comprises single-stranded nucleic acid templates and a second set of nucleic acids comprises at least one set of nucleic acid fragments; and,

(2) elongating, ligating, or both, sequence gaps between the hybridized nucleic acid fragments, to generate one or more substantially full-length chimeric nucleic acid sequences that correspond to the single-stranded nucleic acid templates, thereby recombining the set of nucleic acid fragments.

122. The method of claim **121**, further comprising:

(3) denaturing the one or more substantially full-length chimeric nucleic acid sequences and the single-stranded nucleic acid templates;

(4) separating the at least substantially full-length chimeric nucleic acid sequences from the single-stranded nucleic acid templates by at least one separation technique; and, fragmenting the separated one or more substantially full-length chimeric nucleic acid sequences by nuclease digestion or physical fragmentation to provide chimeric nucleic acid fragments.

123. A method of shuffling polynucleotides, comprising:

initiating a polynucleotide amplification process on overlapping segments of a population of variant polynucleotides, at least one of which variant polynucleotides encodes a dioxygenase polypeptide or a fragment thereof, under conditions

whereby one segment serves as a template for extension of another segment, to generate a population of recombinant polynucleotides; and

selecting or screening a recombinant polynucleotide for a desired property.

124. The method of claim **123**, wherein the overlapping segments are produced by cleavage of the population of variant polynucleotides.

125. The method of claim **123**, wherein the cleavage is by DNaseI digestion.

126. The method of claim **123**, wherein the overlapping segments are produced by chemical synthesis.

127. The method of claim **123**, wherein the overlapping segments are produced by amplification of the population of polynucleotides.

128. The method of claim **123**, wherein the population of variant polynucleotides are allelic variants.

129. The method of claim **123**, wherein the population of variant polynucleotides are species variants.

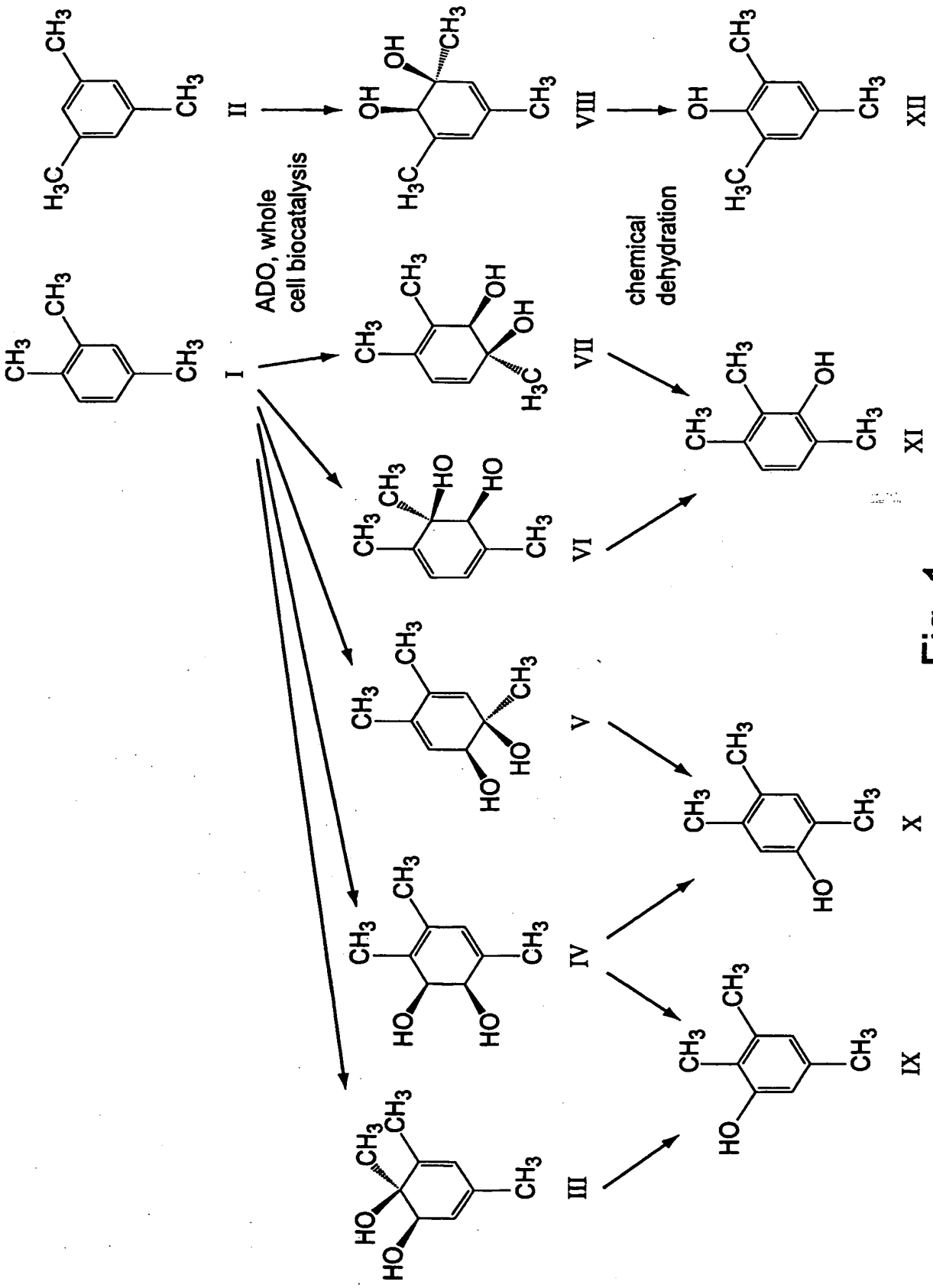


Fig. 1

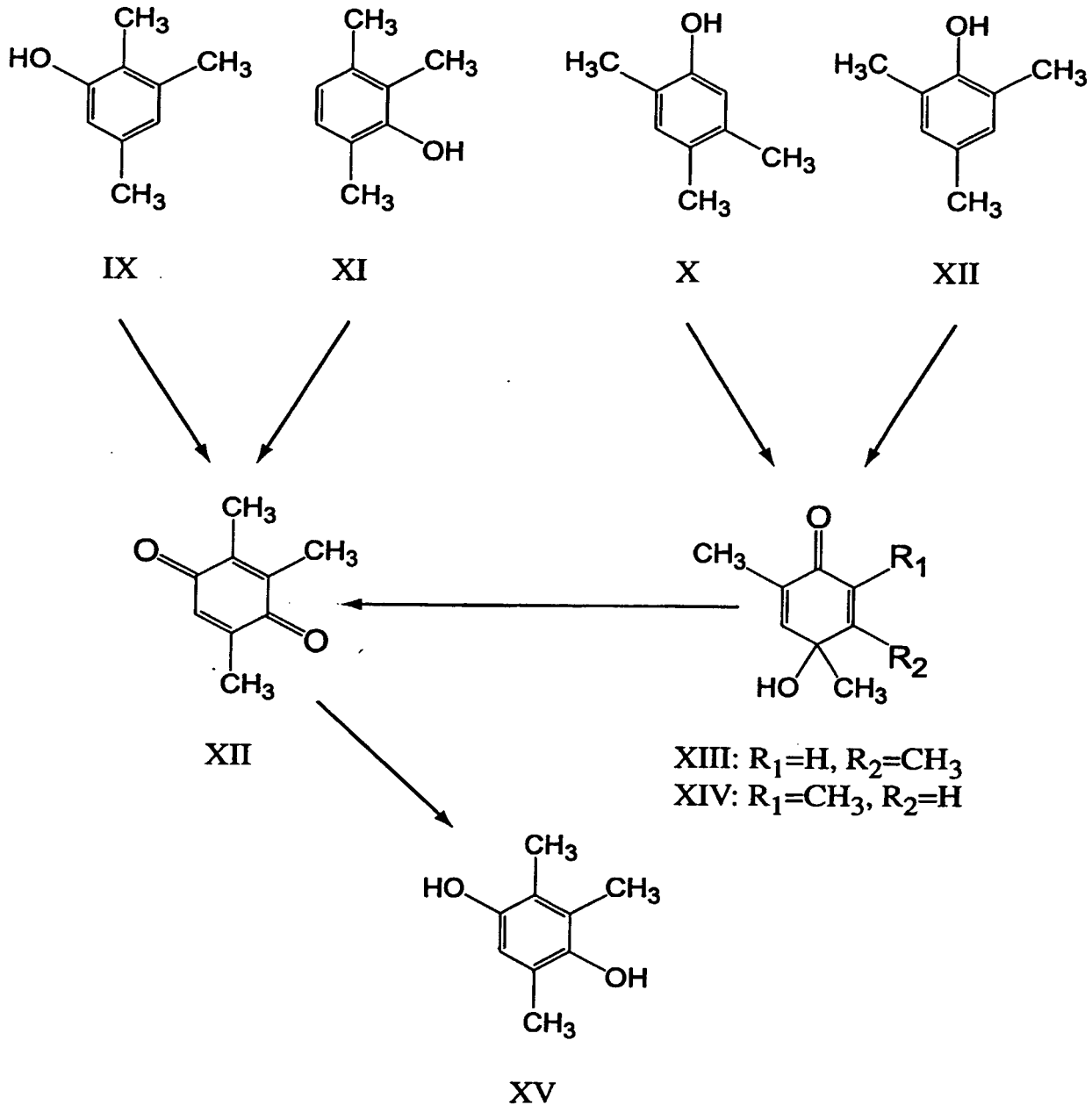


Fig. 2

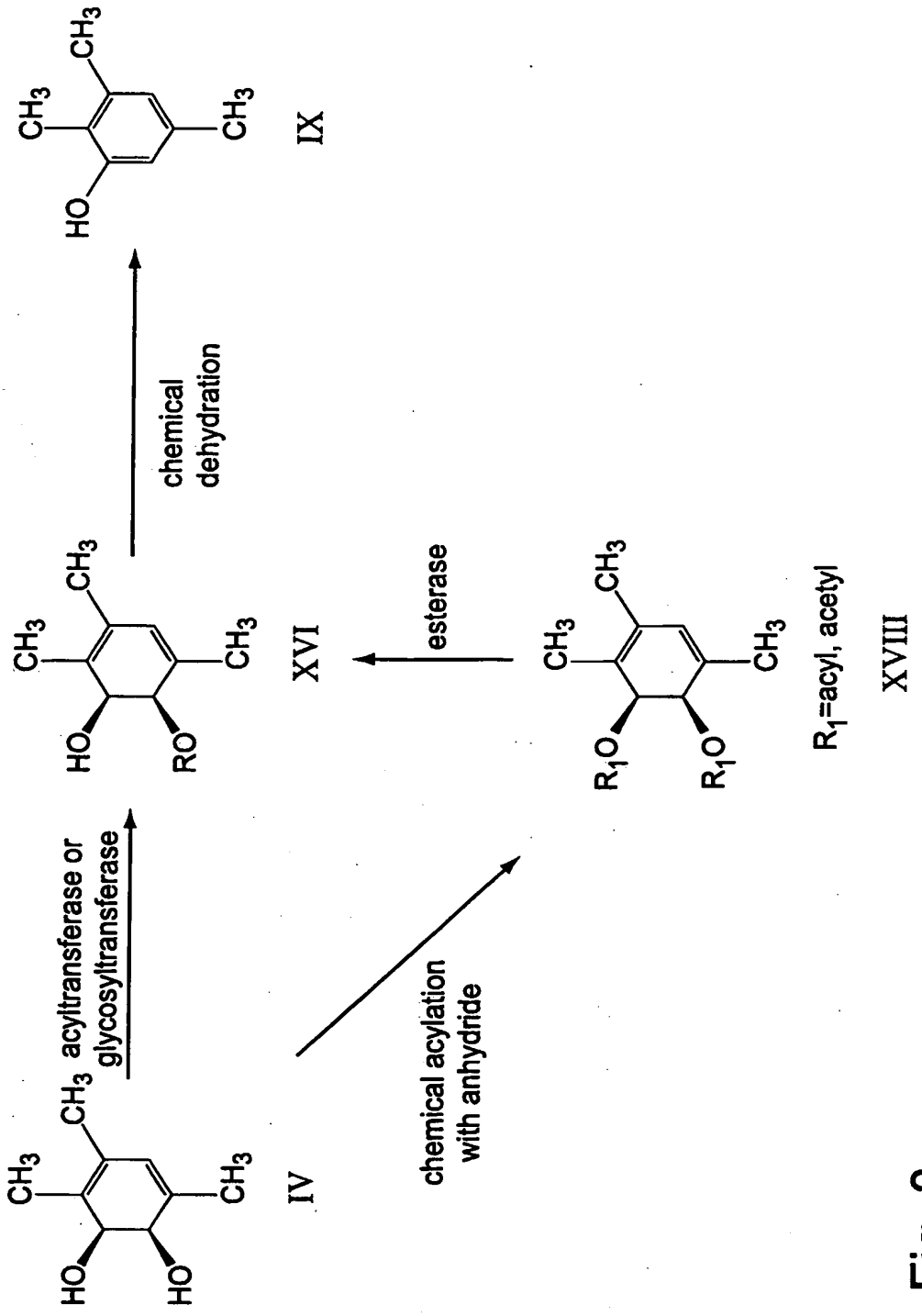


Fig. 3

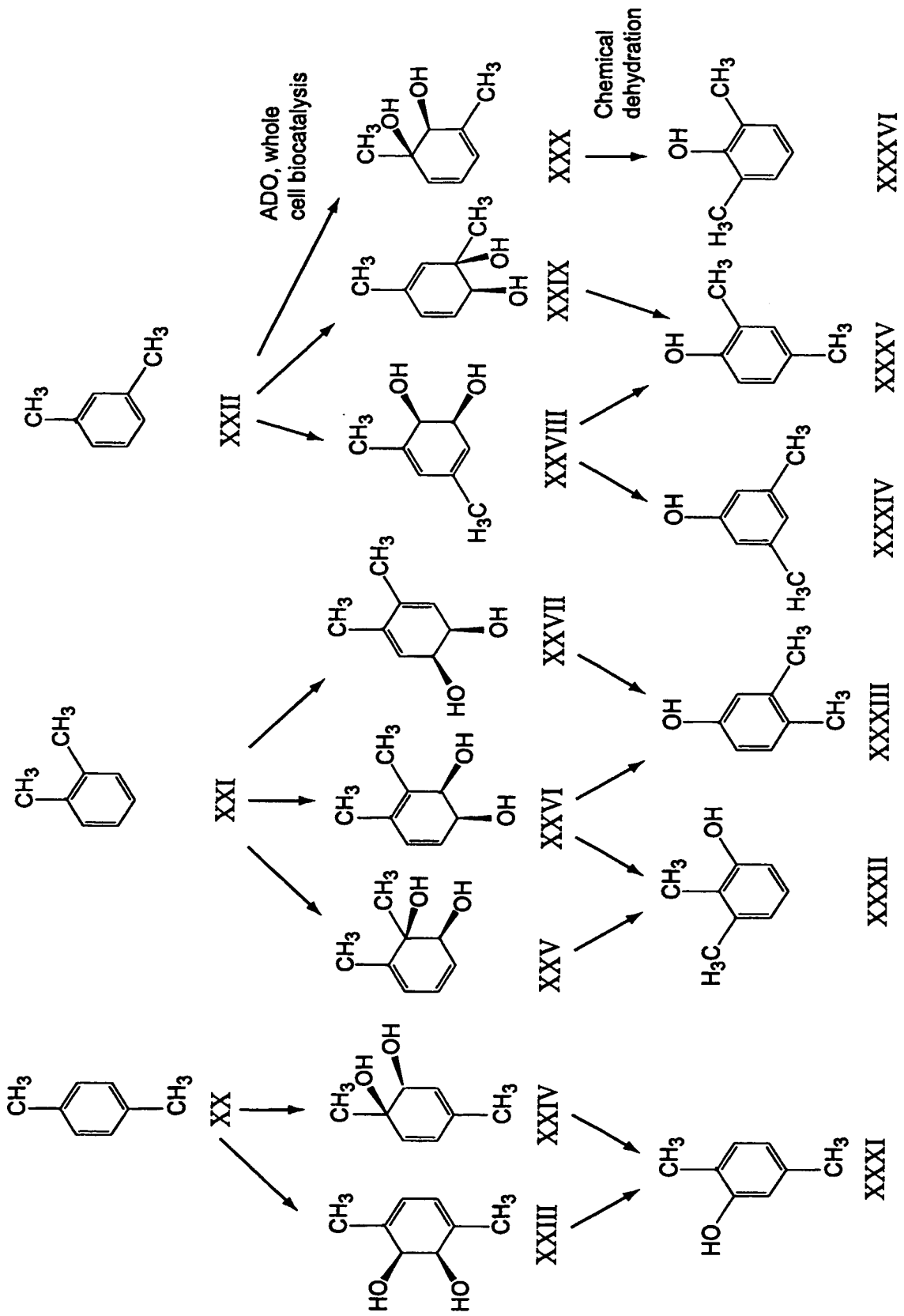


Fig. 4

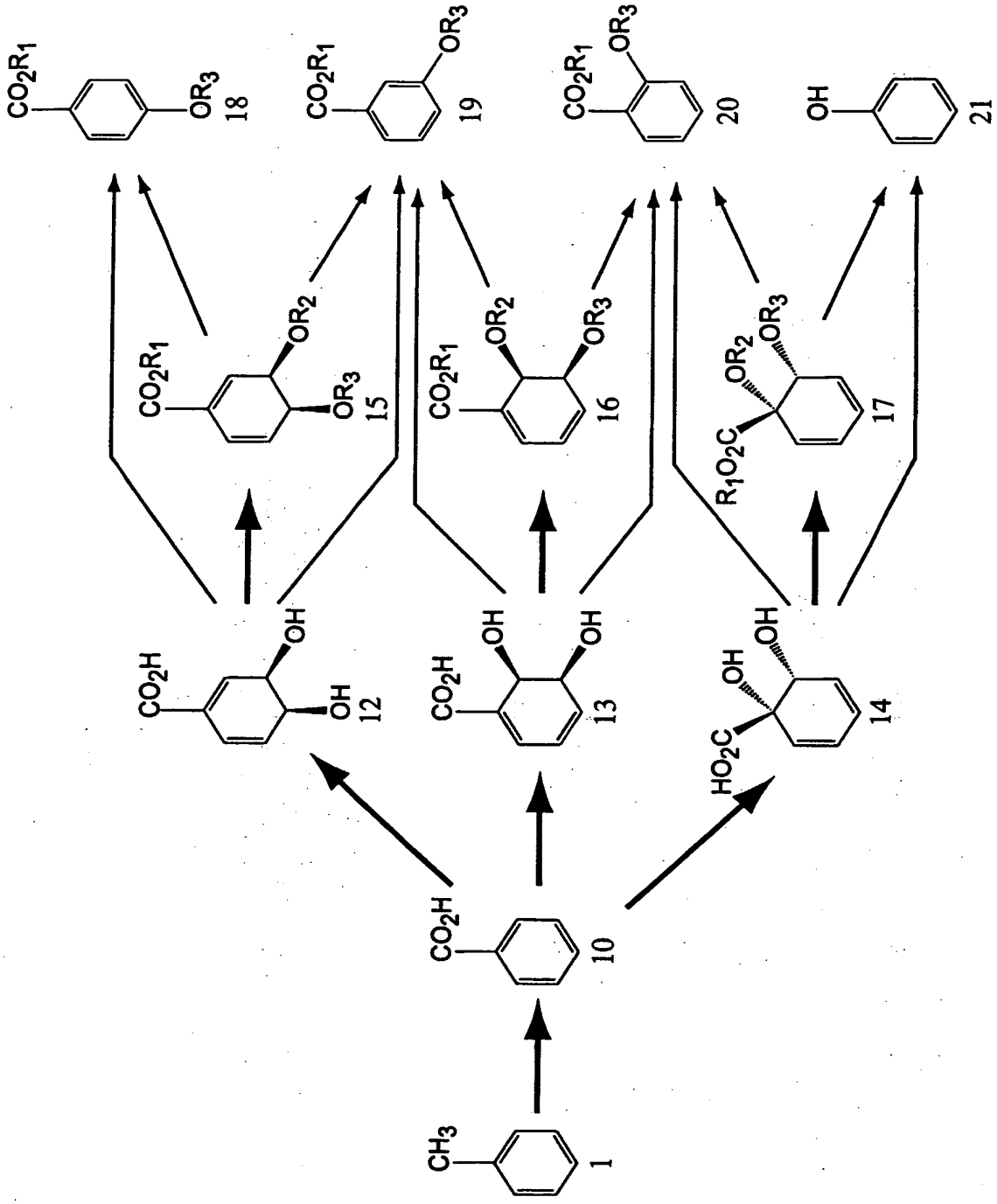


Fig. 5

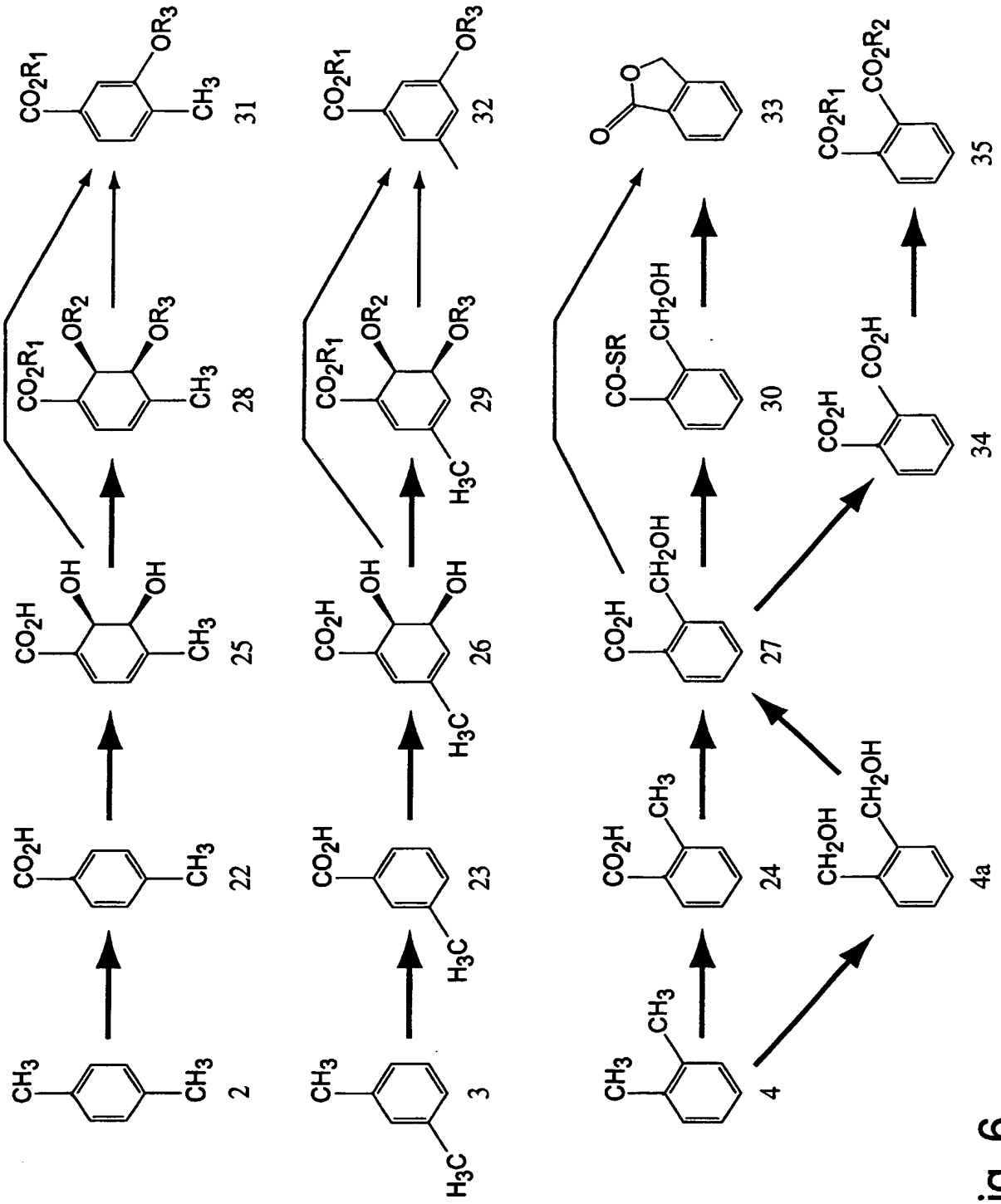


Fig. 6

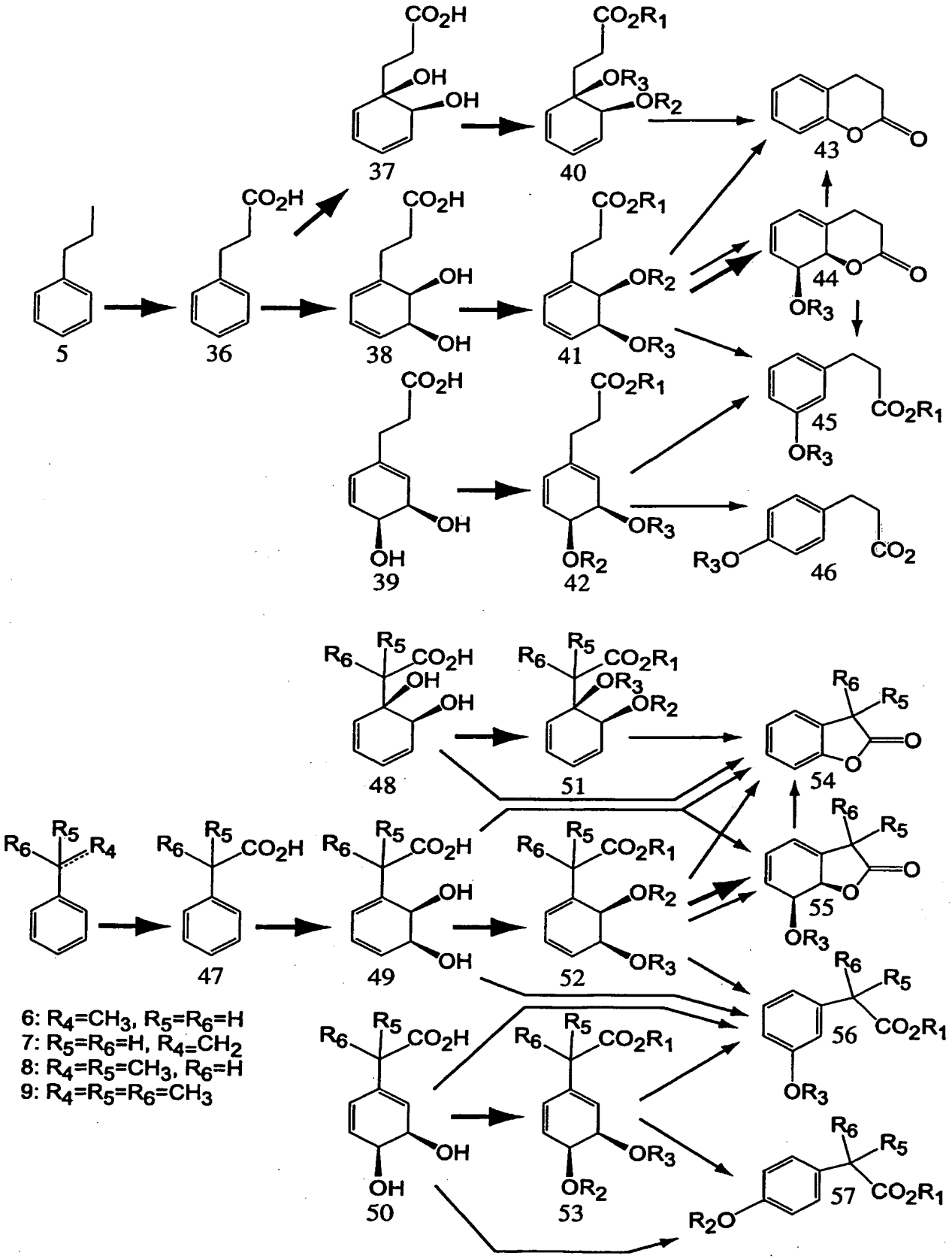


Fig. 7

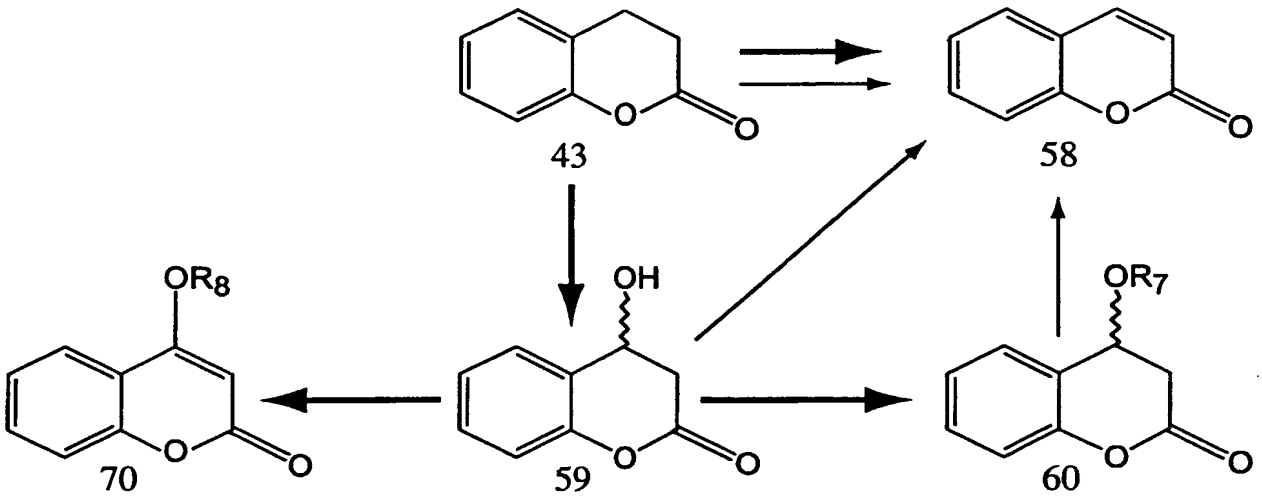


Fig. 8

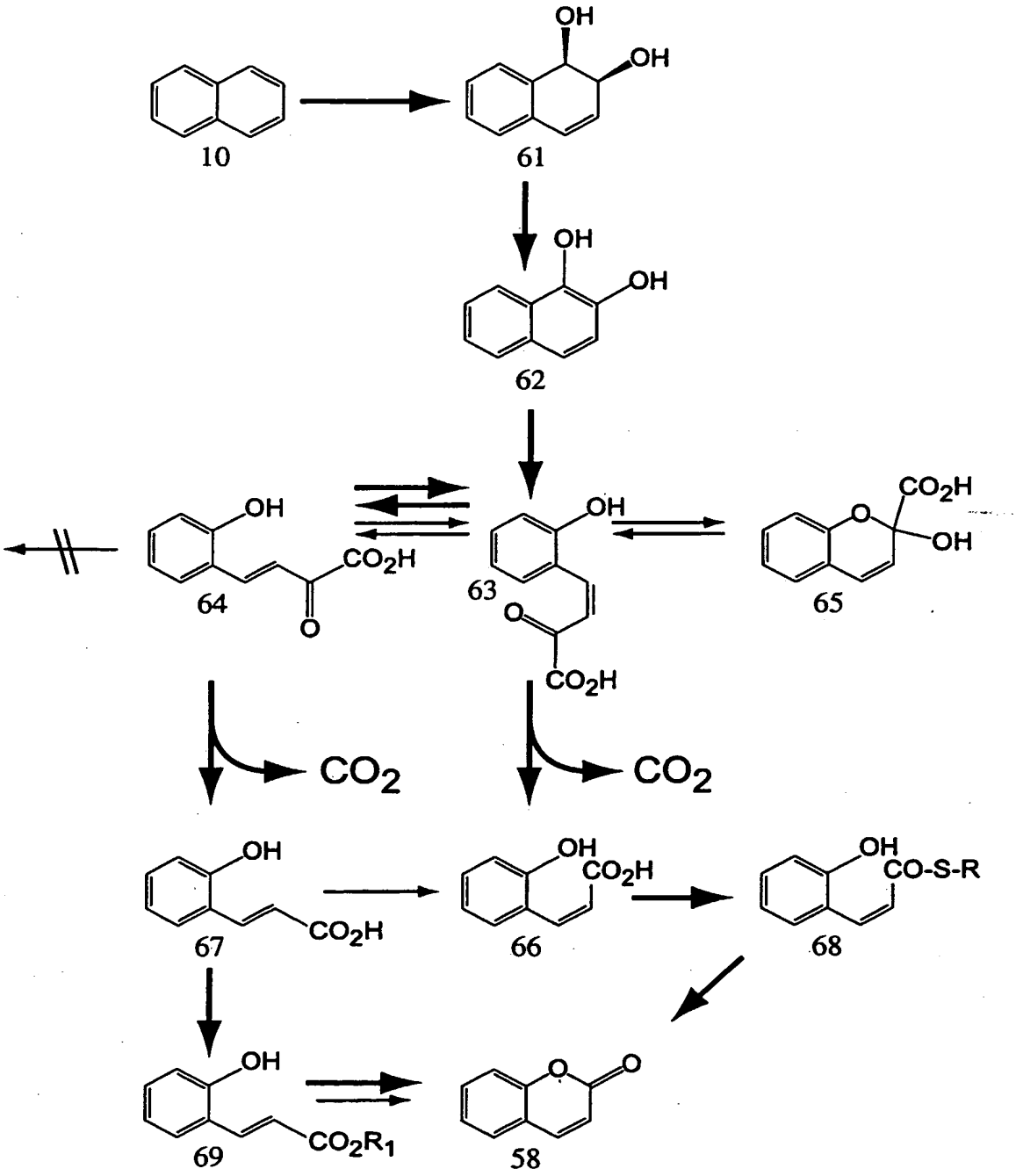


Fig. 9

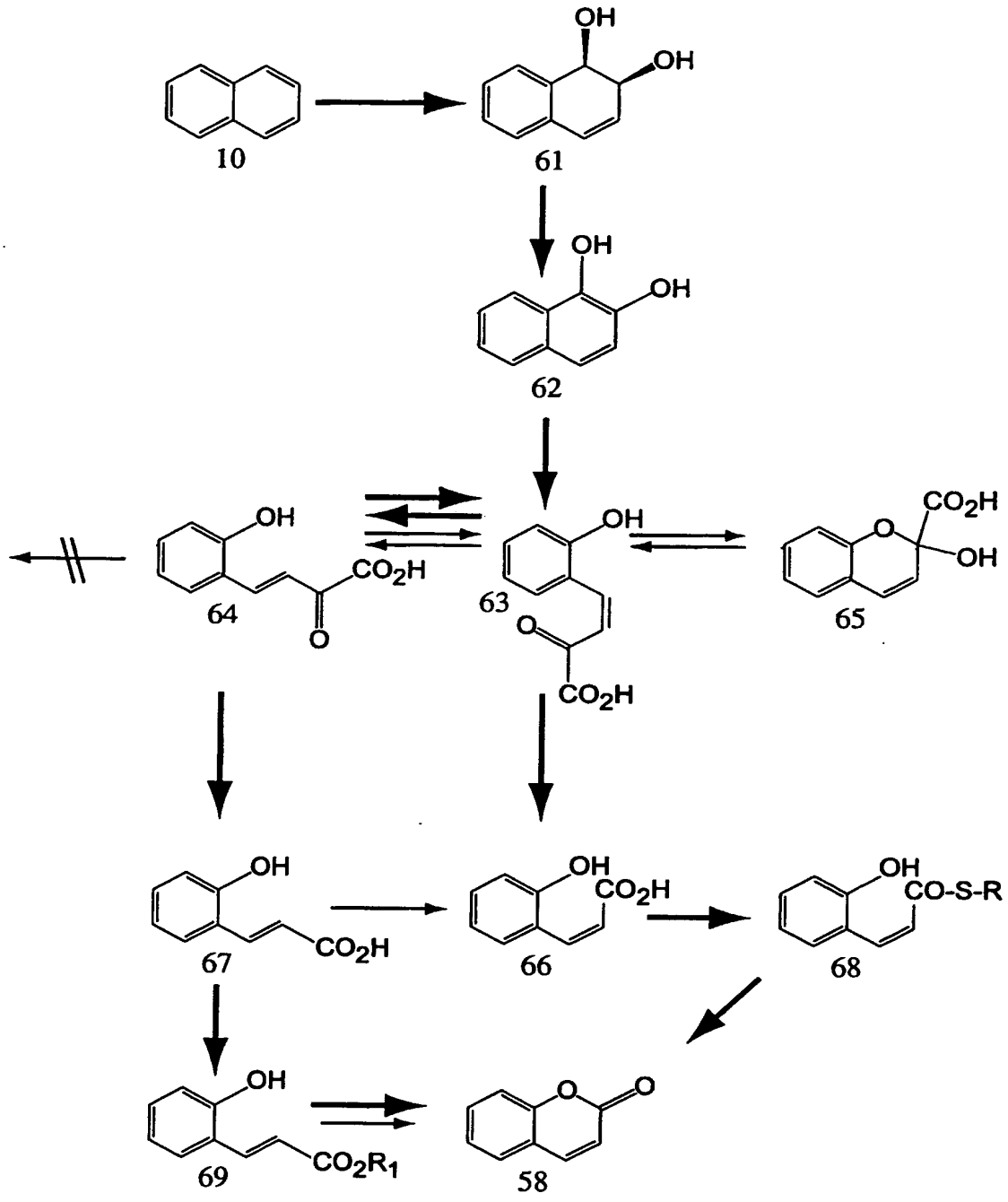


Fig. 10

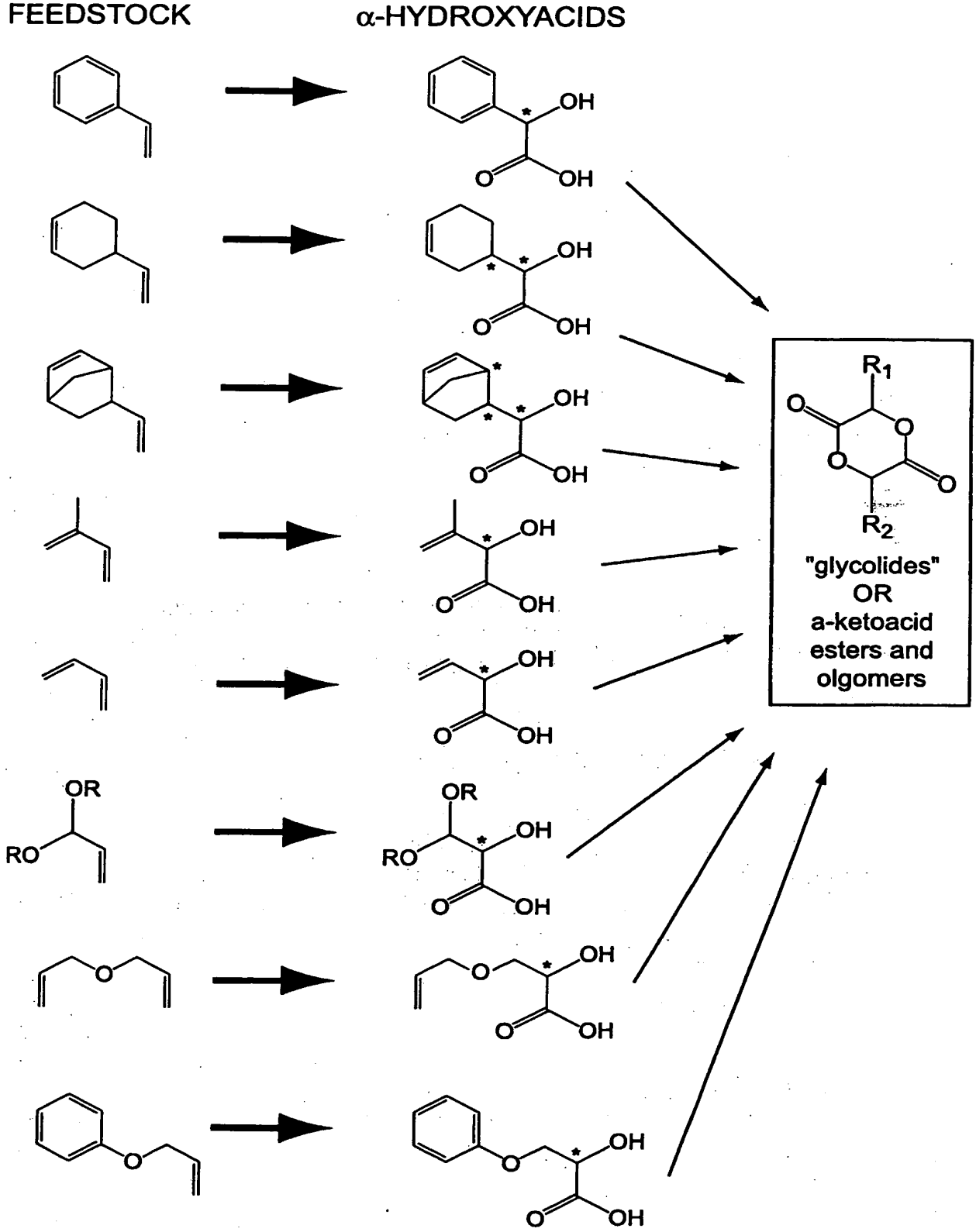


Fig. 11

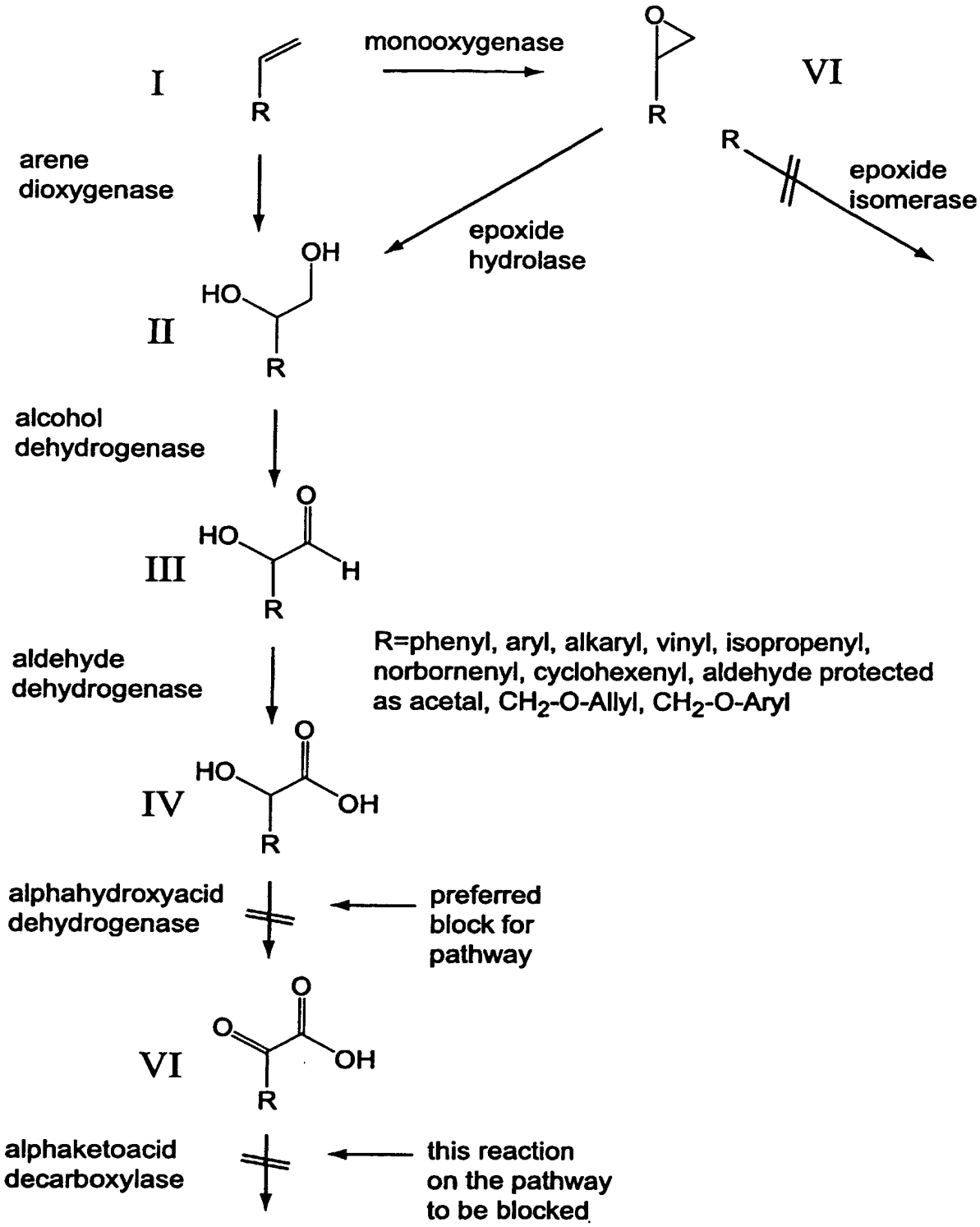


Fig. 12

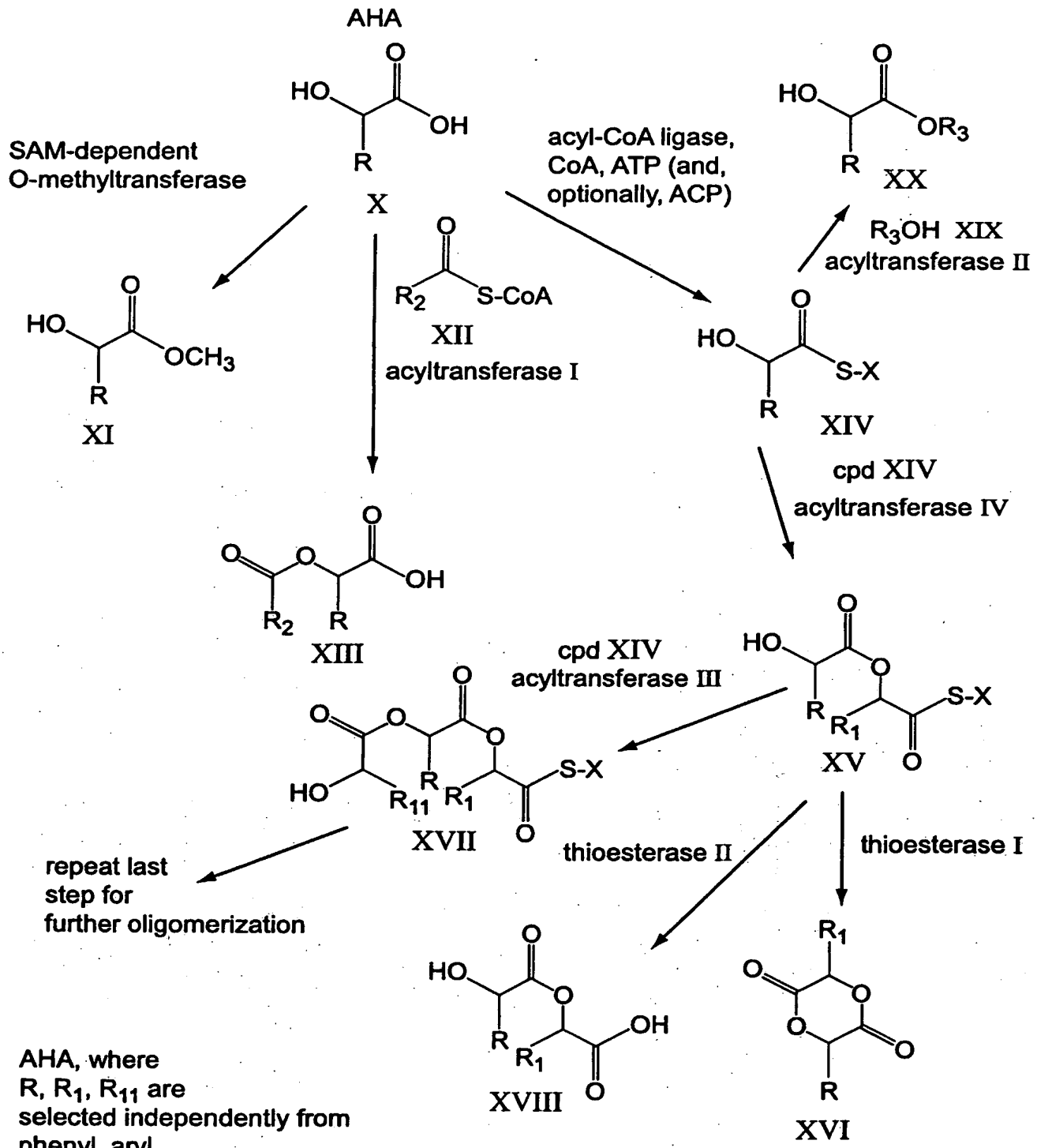
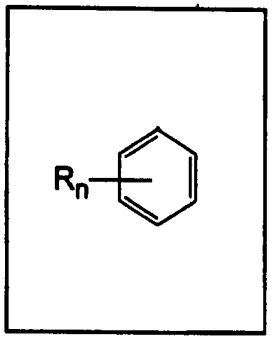
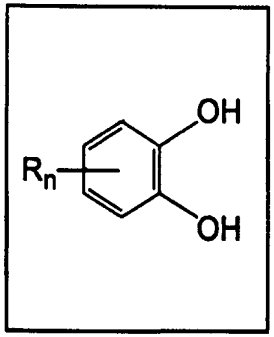


Fig. 13

Sets of and subsets of preferred examples of reactions and substrates	Substrates	Reaction (preferred examples of ADO for use in the reaction and for DNA shuffling to improve characteristics of enzymes for catalysis of preferred reactions)	Products	Types of preferred catalytic properties of ADOs targeted for improvement by DNA suffling	Utility for processes and pathways to make alternative or additional products by subsequent biochemical or chemical reactions (in additions to those formed by ADO)
1. Reactions of ADOs acting as dioxygenases of pi-bonds which are part of benzenoid aromatic rings (reductive dioxygenation of aromatic ring to form *cis*-vicinal diols of conjugated dienes, whether isolatable/stable, or not)					
1.1. Monocyclic aromatic compounds without carboxyl groups attached directly to aromatic ring		all known ADOs		1. Rate of reaction 2. Regio-selectivity, absolute configuration and enantiomeric purity	

n=0-4, any substitution pattern,
Each R is having from 1 to 15 carbon atoms and independently selected from H, alkyl (normal or branched, cycloalkyl, alkenyl, alkynyl, aralkyl, alkoxy-, aryloxy-, alkylthio-, halogen (F, Cl, Br, I) CF_3, CN, NO_2, trimethylsilyl, trimethylgermanyl, trimethylstannyl, alkylamine (CHx)mNR'R", alkyloxy-alkyl or alkyloxyaryl (CHx)m0-R', alkyl-thiaalkyl (CHx)mS-R',
where R' has 1-10 carbon atoms;
m is 1-5 and x is 0-2

R also can contain: protected ketone free or protected as ketal or N-oxime aldehyde, protected as acetal or N-oxime alcohol (primary, secondary or tertiary, protected or not) carboxyl (free or ester or N-containing carboxyl derivative including nitrile or amide, or mono/di N-alkylamide)

Fig. 14

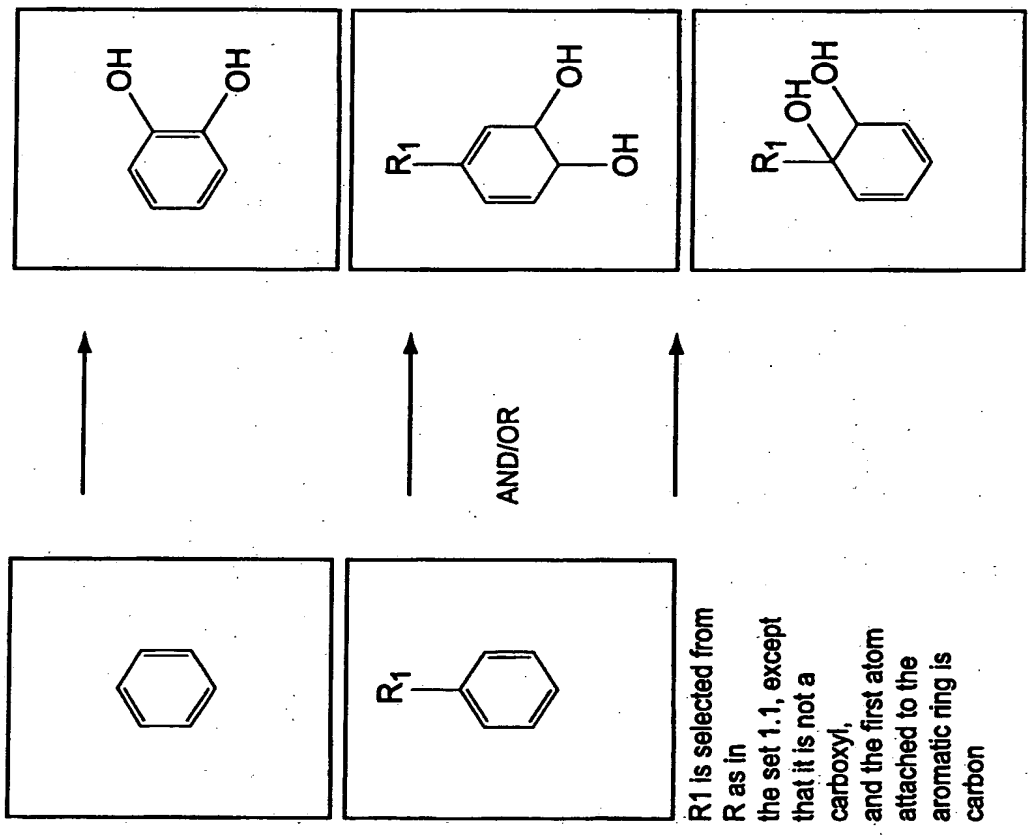


Fig. 14B

1.1.2.1.
benzenes with tethered carboxyl groups
where R, is
-(CH2)m-,
m=1-10
-CH(CH3)-
-CH=CH-
AND/OR
AND/OR
1. rate of reaction
2. regio-selectivity of dihydroxylation absolute stereochemistry and enantiomeric purity of the cis-diol products
intermediates to make hydroxy-aromatic acids (via chemical dehydration of cis-diols) in the biochemical pathway allowing for converting corresponding alkyl-benzenes to the carboxylic substrates prior to dioxygenase reaction
1.1.3. xylenes
1. Rate of oxidation,
2. Regioselectivity of ring dihydroxylation versus methyl group oxidation
3. Specificity towards preferential oxidation of a particular isomer when mixed xylenes are used as substrates.
Various xylenol (dimethyl-phenol) isomers (by dehydration of the cis-diols of their derivatives.
1.1.3.1.
p-xylene
1. Rate of oxidation
2. Specificity when other xylene isomers are present in substrate mixture
2,5-xylenol (via dehydration of the cis-glycol)

Fig. 14C

Fig. 14D

1.1.5.

1,2,4-trimethyl-benzene (pseudo-cumene)

AND/OR

1. Rate of oxidation, in combination with 2. Regioselectivity of dihydroxyla-tion at shown positions versus oxidation of other positions or methyl group(s).

trimethylphenol isomers for vitamin E synthesis (by dehydration of the cis-diols or their derivatives)

AND/OR

1.1.6.

2-halo-p-xylene and other 2-substituted p-xylenes with heteroatom substituents

where R, is halogen (F, Cl, Br, I) or methoxy-

1. Rate of oxidation, in combination with 2. Regioselectivity of dihydroxyla-tion at shown position versus oxidation of other positions or methyl group(s)

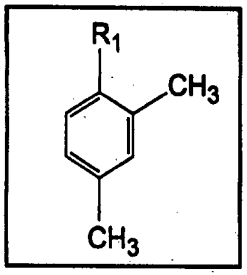
ADOs shuffled to catalyze this easily detectable reaction are suitable for converting 1, 2, 4-trimethylbenzene to angular cis-diols in set #1.1.5

unstable product, decomposes in water to catechol and HR_1

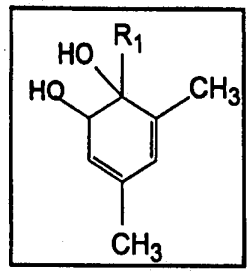
Fig. 14E

1.1.7.

4-halo-m-xylene and other 4-substituted m-xylenes with heteroatom substituents



where R_1 is halogen (F, Cl, Br, I) or methoxy-

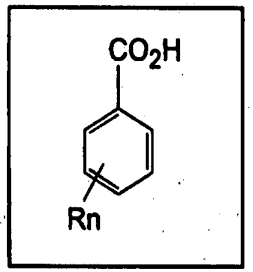


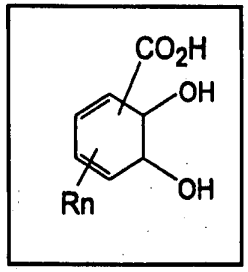
unstable product, decomposes in water to catechol and HR_1

1. Rate of oxidation, in combination with
2. Regioselectivity of dihydroxyla-tion at shown positions versus oxidation of other positions or methyl group(s)

ADOs shuffled to catalyze this easily detectable reaction are suitable for converting 1, 2, 4-trimethylbenzene to angular cis-diols in set #1.1.5

1.2. Monocyclic aromatic compounds having a carboxyl group attached directly to aromatic ring. (benzoic acids)





1. Rate of oxidation
2. Regioselectivity - of ring dihydroxylation at a particular position

3, Specificity for one or more benzoate when other substituted benzoate isomers are present in substrate mixture

intermediates to make hydroxy-aromatic acids (via chemical dehydration of cis-diols) in the biochemical pathway allowing for converting corresponding methyl-benzenes to the benzoates prior to dioxy-genase reaction.

The cis-diols can also be used as building blocks for natural product synthesis and combinatorial libraries of compounds

The hydroxybenzoic acids can be used as monomers for making various aromatic polymers and co-polymers of polyester type.

where n is 0-2 and
R is H, CN, halogen,
$-CF_3$, or substituent having 1-10 carbon atoms and is independently selected from alkyl, cycloalkyl, alkoxy, aryl, aryloxy-, aralkyl-,
alkenyl, acyl,
alkylthio-,
alkyloxy-including the above radicals derivatives having ketone (free or protected as ketal /N-oxime) or aldehyde (protected as acetal or N-oxime),
amine, alkylamine or halide groups

Fig. 14F

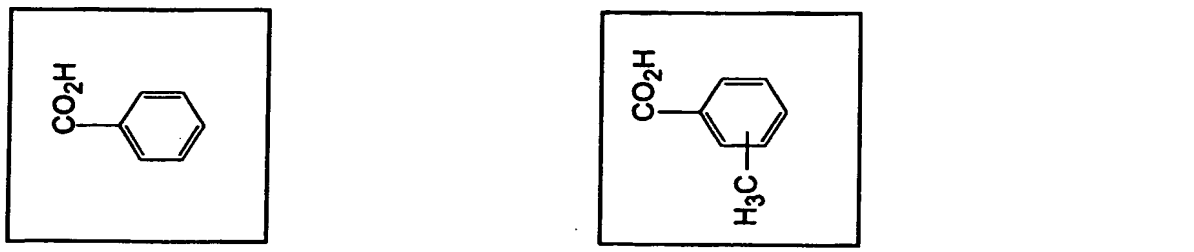
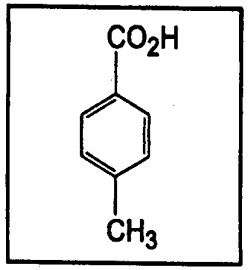
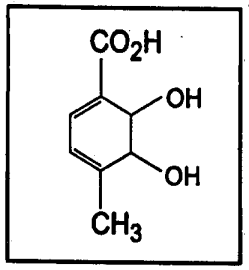


Fig. 14G

1.2.2.1

p-toluic acid





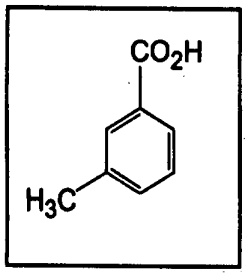
1. Rate of reaction
2. Specificity towards p-toluic acid when a mixture of substituted benzoates is used as substrates

Dehydration of the cis diols yields exclusively 3-hydroxy-p-toluic acid

The diol is preferred intermediate product to make 3-hydroxy-p-toluic acids (via chemical dehydration) in the artificial biochemical pathway allowing for converting p-xylene to p-toluate prior to the dioxygenase reaction.

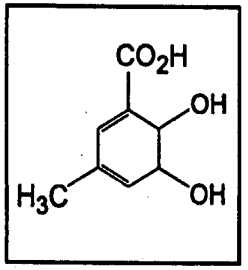
1.2.2.2.

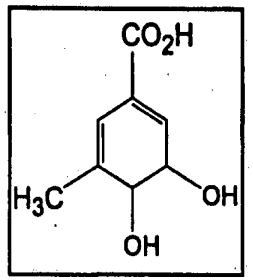
m-toluic acid

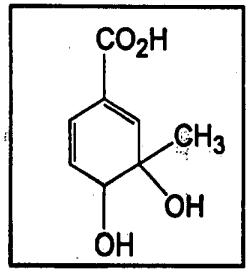


AND/OR

AND/OR







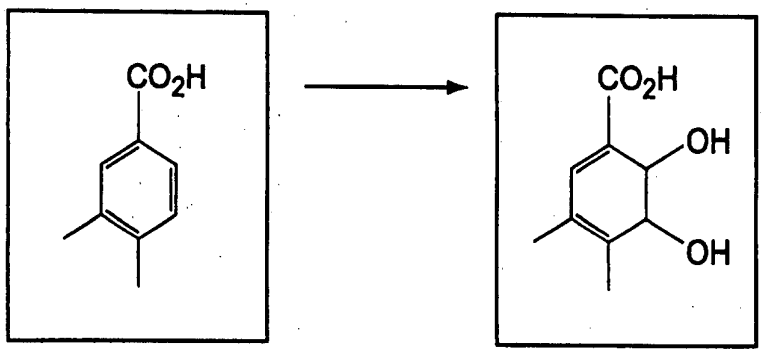
1. Rate of reaction
2. Regioselectivity of ring oxidation at shown positions
3. Specificity towards m-toluic acid when a mixture of substituted benzoates is used as substrate

The diol is preferred intermediate product to make 3- 4-hydroxy-m-toluic acids (via chemical dehydration) in the artificial biochemical pathway allowing for converting m-xylene to m-toluate prior to the dioxygenase reaction.

Fig. 14H

Fig. 14I

1.2.3.1.
3, 4-dimethyl-benzoic acid



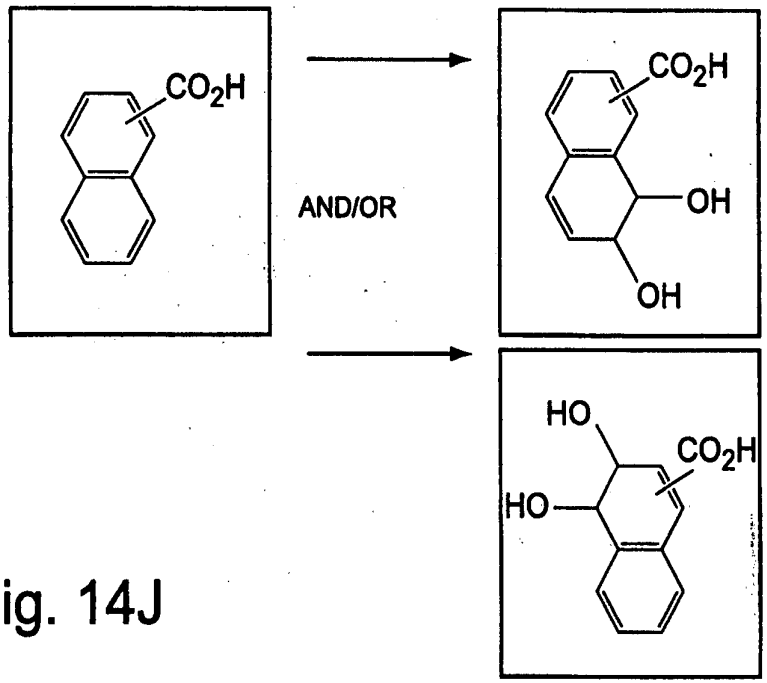
1. Rate of reaction
2. Regioselectivity of ring oxidation at shown positions
3. Specificity towards 3, 4-dimethyl-benzoic acid when a mixture of dimethyl-substituted benzoates is used as substrate

Dehydration of the cis-diols yields exclusively 3-hydroxy-4, 5-di methylbenzoic acid,

The diols are preferred intermediate product to make 3-hydroxy-4, 5-dimethylbenzoic acid (via chemical dehydration) in the artificial biochemical pathway allowing for converting 1, 2, 4-trimethylbenzene to 3, 4-dimethylbenzoic acid prior to the deoxygenase reaction.

1.3. Aromatic compounds having more than one aromatic ring, whether benzenoid or heterocyclic

1.3.1.
naphthoic acids



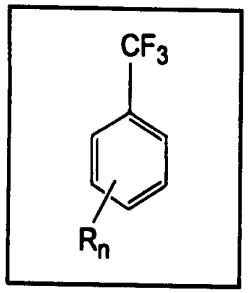
1. Rate of dihydroxylation
2. Regioselectivity for oxidation of various p-bonds, including oxidation of ring possessing the carboxyl and ring which does not have the carboxyl.
3. Absolute configuration and enantiomeric purity of the cis-diols

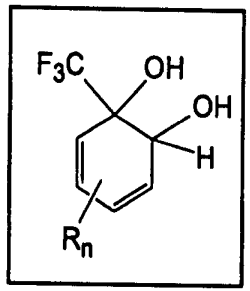
The cis-diols can be readily dehydrated to a variety of hydroxynapthoic acids of various substitution patterns. These compounds are useful monomers for making aromatic polyester polymers and co-polymers

The dioxygenases are particularly useful as a part of artificial biochemical pathway allowing for converting 1- and 2-methylnaphthalenes to the corresponding naphthoic acids prior to the dihydroxylation reaction of the aromatic ring

Fig. 14J

1.4. Aromatic compounds having at least one trifluoro-methyl(-CF_3) substituent attached directly to a benzenoid aromatic ring





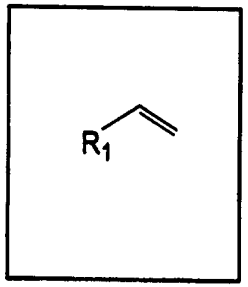
where n is 0-2 and each R is independently selected from COOH, halogen, CF_3, CN, $CONH_2$ or alkoxyl

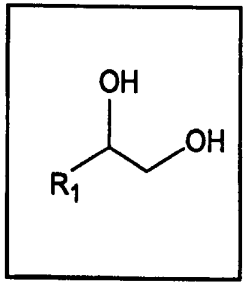
1. Rate of reaction
2. Regioselectivity of ring dihydroxylation at the p-bond bearing CF_3 substituent
3. Absolute configuration and enantiomeric purity of the diol

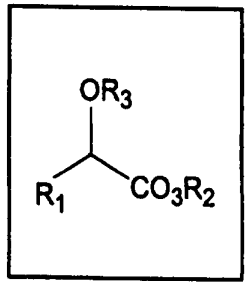
Compounds are useful for making substituted p-poly-phenylene polymers, as well as building blocks for combinatorial libraries of compounds

2. Reactions of ADOs acting as dioxygenases of pi-bonds which are not part of benzenoid aromatic rings, whether conjugated or not with other pi-bonds (dihydroxylation to vicinal glycols)

2.1.
Alpha-olefins







alpha olefins, where
R_1 is selected from C_1-C_{30} alkyls (normal and branched), cycloalkyl, cycloalkenyl, aryl, aralkyl, pyridyl, furanyl, thienyl, alkoxy, aryloxy, aralkyloxy, alkylamino, dialkyl-amino, arylamino-

and alpha-amino acids, acrylic acids, a-ketoacids and corresponding esters

Fig. 14K

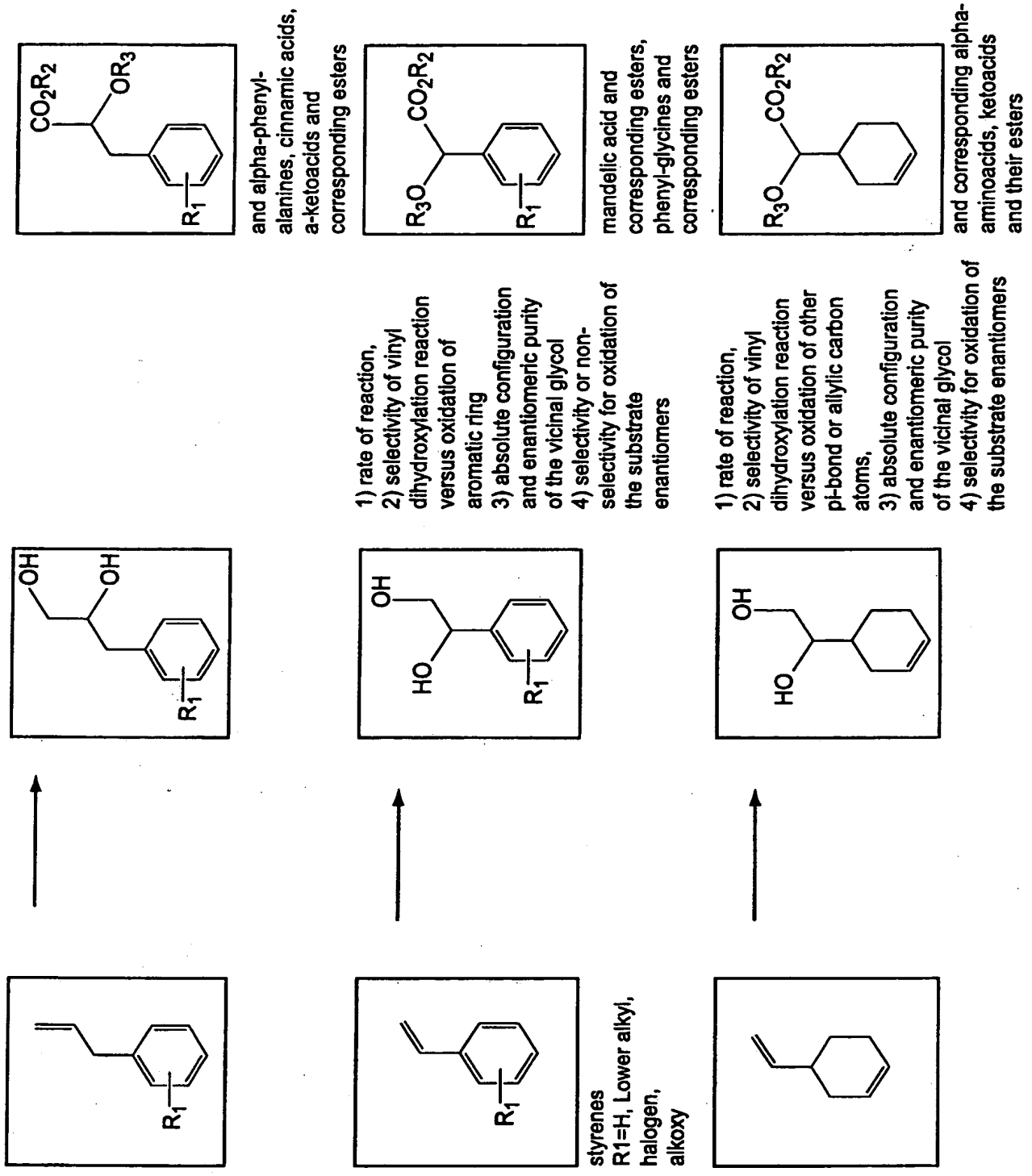


Fig. 14L

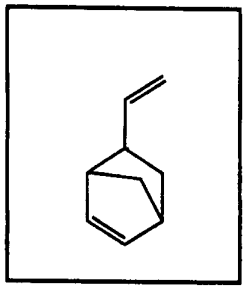
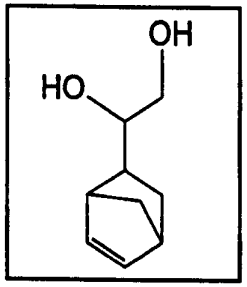
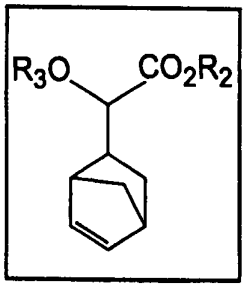
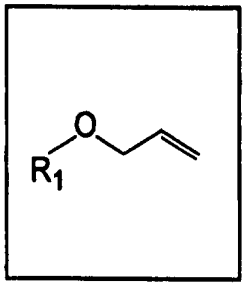
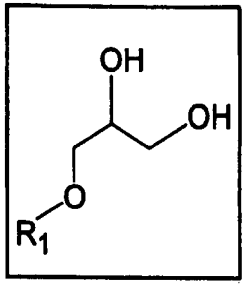
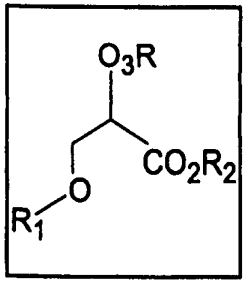
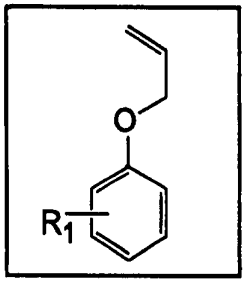
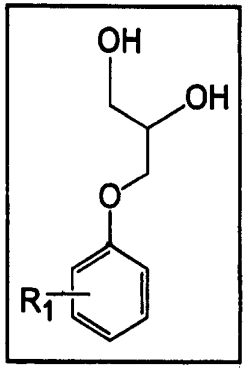
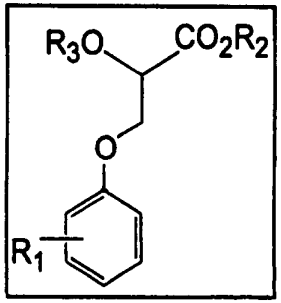
2.5. 2-vinyl-norborn-5-ene			1) rate or reaction, 2) selectivity of vinyl dihydroxylation reaction versus oxidation of other pi-bond, 3) absolute configuration and enantiomeric purity of the vicinal glycol 4) selectivity or non-selectivity for oxidation of substrate enantiomers and/or syn-anti vinyl stereoisomers.	 and correspond-ing alpha-animoacids, ketoacids and their esters
2.5. Allyl ethers	 Where R1 is alkyl, allyl, phenyl, aryl, alkaryl alkylphenyl		1) rate or reaction, 2) selectivity of allylic pi-bond dihydroxylation reaction versus oxidation of other pi-bonds, if present in the substrate stucture 3) absolute configuration and enantiomeric purity of the vicinal glycol	 and alpha-amino acids, a-ketoacids and corresponding esters
2.5.1 Allyl phenyl ethers	 Where R1 is lower alkyl		1) rate or reaction, 2) selectivity of allylic pi-bond dihydroxylation reaction versus aromatic ring oxidation 3) absolute configuration and enantiomeric purity of the vicinal glycol	 and alpha-amino acids, acrylic acids, a-ketoacids and corresponding esters

Fig. 14M

2.6. cyclic alkenes

Where R1 and R2 are carbon radicals of various length forming at least one cyclic ring.

1) rate or reaction,
2) selectivity of dihydroxylation reaction versus competing allylic mono-oxygenation
3) absolute configuration of cis-glycol, where applicable

mono and diesters of cyclic glycols, ketones, acyloins (keto-alcohols), diacids, dialdehydes, omega-ketoacids, omega-hydroxy-acids and corresponding lactones

2.6.1.

cyclohexene

1) rate of dihydroxylation
2) selectivity of pi-bond dihydroxylation versus allylic mono-hydroxylation

mono- and diesters of cyclohexane-1, 2-diol, adipoin, adipic acid, adipic semialdehyde, epsilon-hydroxy-caproic acid, caprolactone, epsilon-amino-caproic acid, caprolactam, cyclohexa-1, 3-diene (by dehydration reaction from cis-glycol)

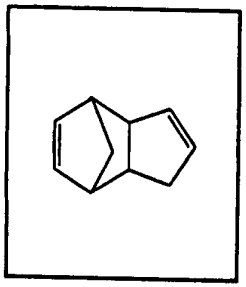
2.7. Indene

1) rate of dihydroxylation
2) selectivity of p-bond dihydroxylation versus benzylic monohydroxylation and aromatic ring oxidation
3) absolute configuration and enantiomeric purity of the cis-glycol

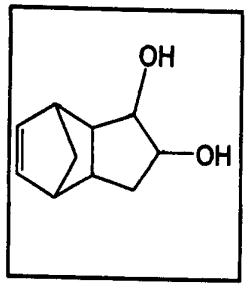
Pharmaceutical intermediate for drug synthesis (e.g. indinavir)

Fig. 14N

2.8.
Dicyclopentadiene



Dicyclopentadiene

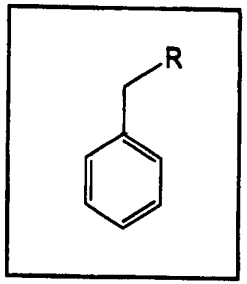


1) rate or reaction,
2) selectivity of dihydroxylation reaction versus oxidation of alternative pi-bond,
3)absolute configuration and enantiomeric purity of the vicinal glycol
4) selectivity or non-selectivity for oxidation of substrate enantiomers and endo/exo stereoisomers

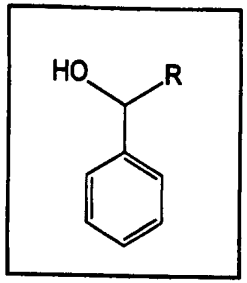
cyclopentenes with chiral hydroxyl centers - pharmaceutical inter-mediates
(the glycol derivative with protected hydroxyls can be subjected to retro-Diels-Alders reaction to produce cyclopentadiene and desired oxygenated cyclopentene derivative)

3. Reactions of ADOs acting as monooxygenases at sp3 carbon atoms adjacent to aromatic rings, whether benzenoid or heterocyclic (monohydroxylation of methyls or benzylic methylene groups to benzylic alcohols)

3.1. alkyl-benzenes



where R is n-alkyl radical having 1-20 carbon atoms

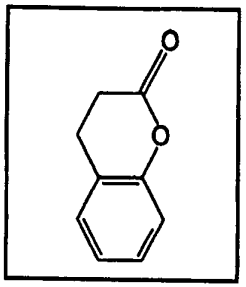


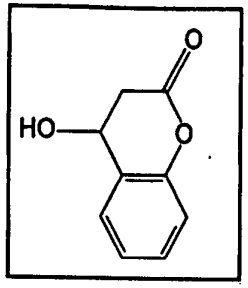
1. Rate of reaction
2. Selectivity of monohydroxylation of benzylic methylene group versus ring oxidation and desaturation
3. Absolute configuration of the product and enantiomeric purity.

Reaction is particularly useful to make phenethyl alcohol (and styrene by chemical dehydration of the later).

Secondary alcohols can be further in-vivo converted to corresponding ketones (with recruitment of appropriate genes encoding alcohol dehydrogenases), from which acetophenone is particularly useful
Reaction is useful to convert dihydro-coumarin to coumarin by acid-catalyzed dehydration of the hydroxylated product
Dihydrocoumarin can in turn be prepared from n-propylbenzene by reactions and process described in the set 1.1.2.1.

3.2.
3, 4-dihydro-coumarin

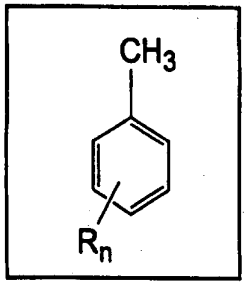


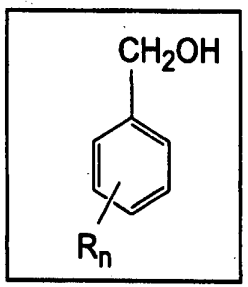


1. Rate of reaction
2. Selectivity for monohydroxylation of benzylic methylene group versus ring oxidation

Fig. 14O

3.3.
methyl-benzenes





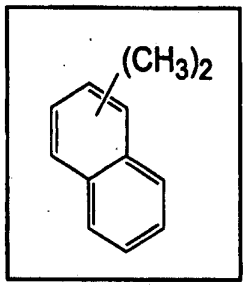
where R is lower alkyl with 1-5 carbon atoms and n is 0-3

1. Rate of reaction
2. Selectivity for oxidaiton of one or more methyl-benzenes from a mixture of methyl-benzenes
3. Selectivity for methyl group oxidation versus ring oxidation and selectivity of oxidation of non-equivalent methyl groups

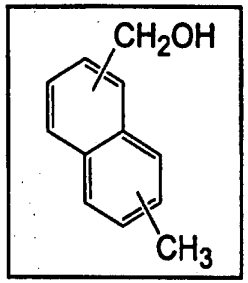
Reaction is useful to achieve catalytic separations of various methylbenzenes by selective oxidation of one or more substrates.

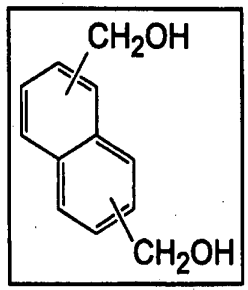
The benzyl alcohols can be further oxidized to corresponding carboxylic acids with provision of appropriate dehydrogenase genes in the same organism which expresses the methyl group oxidizing activity.

3.4.
dimethylnaphthalenes



AND/OR





1. Rate of reaction
2. Selectivity for oxidation of a particular dimethylnaphthalene isomer or a group thereof
3. Ability to oxidize both methyl groups in stepwise fashion
4. Selectivity for oxidation of methyl groups versus aromatic ring oxidation

The utility of the enzymes optimized for methyl group oxidation reaction on benzenes can also be extended to oxidizing methyls of 1- and 2-mono-methylnaphthalenes
Reaction is useful to make naphthalene dicarboxylic acids (with concomitant recruitment of appropriate alcohol- and aldehyde-dehydrogenase genes, and to achieve reactive separation of dimethylnaphthalene isomers.
Most preferred dimethylnaphthalene isomer for selective oxidation is 2, 6-isomer, and such reaction is useful for making naphthalene 2,6-dicarboxylic acid, important polyester monomer.

4. Reactions of ADOs acting as monooxygenases at sp3 carbon atoms adjacent to olefinic pi-bonds (allylic monohydroxylation to allylic alcohols)

Fig. 14P

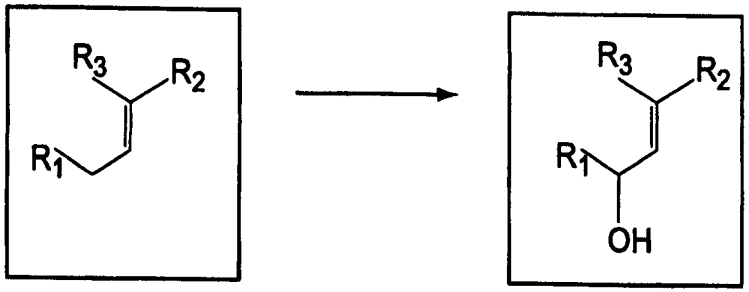
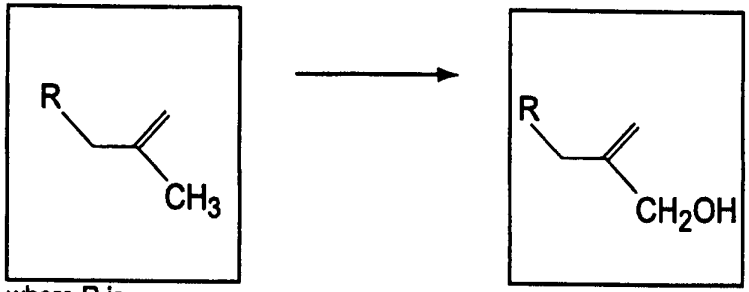
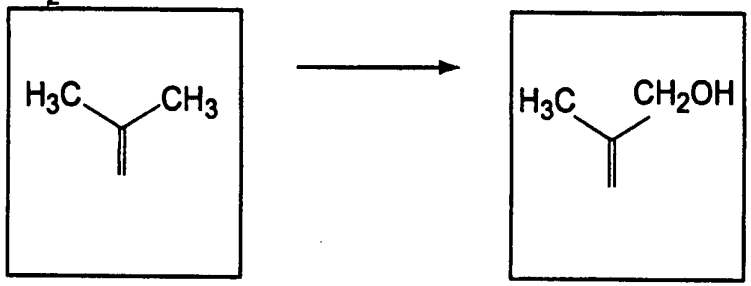
4.1. cyclic unsaturated mono-terpenes		1. Rate of reaction 2. Selectivity of hydroxylation for one or more allylic methyls or methylenes 3. Absolute configuration and enantiomeric purity of the alcohol product 4. Selectivity for oxidation of one terpenoid compound (or its enantiomer) in the presence of other monoterpenoid compounds	The reactions are useful to make high value oxygenated derivatives of monoterpenoids, particularly in the context of engineered metabolic pathways with recruitment of dehydrogenases and other enzymes. The preferred products include carvones, (iso)piperidenones (intermediates in the metabolic pathways leading to menthol) from limonenes, verbenone and myrtenol from pinenes.
	R1, R2, R3 - H or lower alkyls, alkenyls with total number of carbons in the molecule is 10 to represent limonene (dipentene), beta-phellandrene a- and b-pinene, each can be racemic or substantially enantiomerically pure, and impure chemical state or mixture		
4.2. isopentenyl compounds		1. Rate of reaction 2. Selectivity for methyl group oxidation versus other oxidations	Reaction is useful to engineer metabolic pathways for making itaconic acid and methylene-butyrolactones by fermentation
	where R is CH_2OH or CO_2H		
4.3. isobutylene (2-methyl-propene)		1. Rate of reaction	Reaction is useful for bioprocess to make methacrylic acid from isobutylene, with recruitment of suitable alcohol and aldehyde dehydrogenase genes

Fig. 14Q

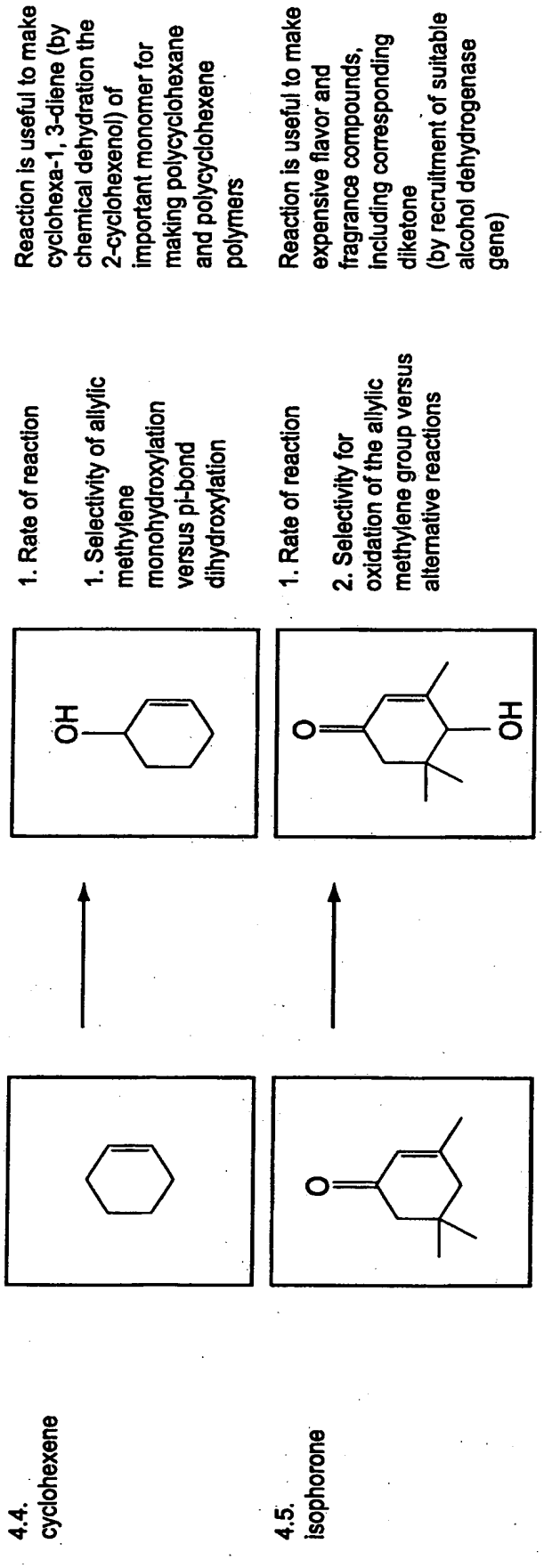


Fig. 14R

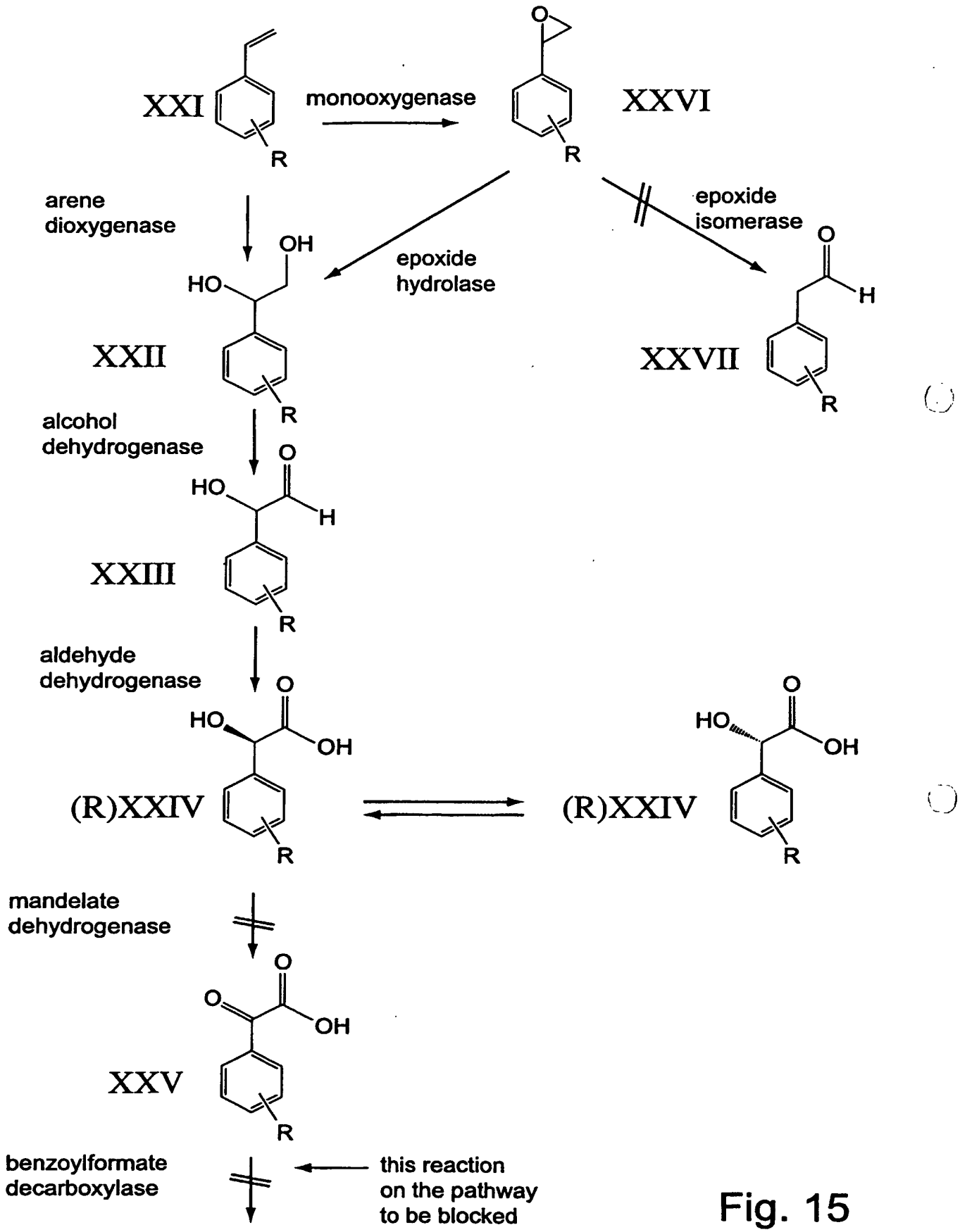


Fig. 15

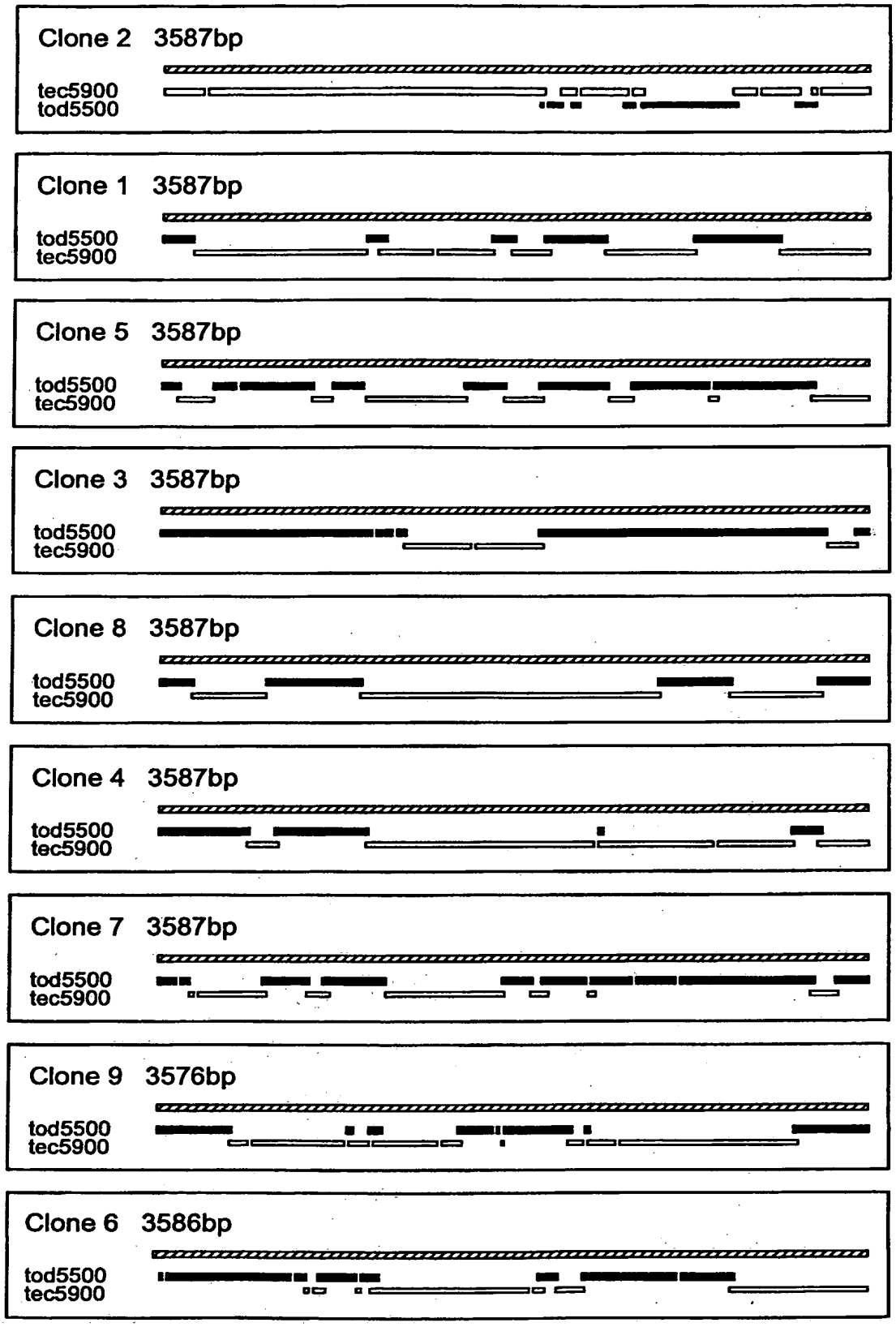


Fig. 16

INTERNATIONAL SEARCH REPORT

Inter nal Application No
PCT/US 00/22038

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 C12N9/02 C12N15/52

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 7 C12N

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)
EPO-Internal, WPI Data, PAJ

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	WO 97 35966 A (MAXYGEN INC ;MINSHULL JEREMY (US); STEMMER WILLEM P C (US)) 2 October 1997 (1997-10-02) page 12, line 9 - line 20 page 22, line 32 -page 23, line 9; claims 1-30	1-129
A	WO 98 27230 A (MAXYGEN INC ;PATTEN PHILLIP A (US); STEMMER WILLEM P C (US)) 25 June 1998 (1998-06-25) the whole document	1-129

☐ Further documents are listed in the continuation of box C.

☒ Patent family members are listed in annex.

° Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
14 December 2000	21/12/2000

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL – 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Sprinks, M

Form PCT/ISA/210 (second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT

Information on patent family members

Inter nal Application No
PCT/US 00/22038

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 9735966 A	02-10-1997	US 6117679 A	12-09-2000
		US 5837458 A	17-11-1998
		AU 2337797 A	17-10-1997
		AU 2542697 A	17-10-1997
		CA 2247930 A	02-10-1997
		EP 0932670 A	04-08-1999
		EP 0906418 A	07-04-1999
		JP 2000507444 T	20-06-2000
		WO 9735957 A	02-10-1997
		US 6096548 A	01-08-2000
		AU 713952 B	16-12-1999
		AU 1087397 A	19-06-1997
		CA 2239099 A	05-06-1997
		EP 0876509 A	11-11-1998
		EP 0911396 A	28-04-1999
		JP 2000500981 T	02-02-2000
		WO 9720078 A	05-06-1997
WO 9827230 A	25-06-1998	AU 5729298 A	15-07-1998
		EP 0946755 A	06-10-1999